1978

Comparison of three methods for prediction: the least square method, ridge regression, and equal weighting

Tetsuro Motoyama
Iowa State University

Comparison of three methods for prediction: The least square method, ridge regression, and equal weighting

by

Tetsuro Motoyama

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of
The Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Departments: Psychology
Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Departments

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa

1978

TABLE OF CONTENTS

CHAPTER I. INTRODUCTION

Application of multiple regression by least square for predicting psychological and educational criteria faces serious problems caused by violation of assumptions and the nature of the variables involved. Yet, researchers in the educational and psychological fields have been using the weights obtained by least square as the "best" weighting method for prediction.

In psychological and educational research and practice, it has been of interest to predict an individual's behavior on a criterion, such as academic success in the form of GPA, from his behavior or measures of behavior on one or more predictor variables, such as aptitude tests. Given the task of predicting behavior on some criterion, the researcher faces several questions. Perloff (1951) summarized these questions:

> (1) What is a relevant criterion? Precisely, what am I seeking to predict? (2) What predictors (tests) shall I use to satisfactorily predict this criterion? (3) Once the predictors are selected how shall I optimally weight each one so that the best estimate (highest and most accurate) of the correlation between the weighted composite and the criterion will be forthcoming? (4) What size N (number of individuals) shall I use in my validation-sample, i.e., a hopefully

representative sample of the population for which prediction is ultimately desired?

The last question suggested by Perloff pointed out that it is necessary to obtain predictor and criterion scores on an initial sample in order to obtain weights for the predictor variables. Without this initial sample, it is not possible to estimate the relationship between predictor variables and a criterion variable, and the prediction problem is reduced to pure guessing or judgments.

The data from the initial sample provide the investigator with the information necessary to estimate the weights for the predictor variables. Once the weights are obtained through some method, the same weights are to be applied to the predictor variables of the subsequent samples of people where the criterion information is not available. Therefore, much of the future prediction depends upon the nature of the initial sample and the particular method employed to determine the weights.

The purpose of this research was to investigate empirically the effectiveness of three weighting methods, namely least square, ridge regression and equal weighting, to predict the criterion variables in future samples.

Questions (1) and (2) by Perloff (1951) were not dealt with in the present research even though they are important and interesting in themselves. It was assumed that the

criterion are well-defined and measurable, and the predictor variables are either given or selected through some method.

Finally, it was the intention of the present research to deal with a model closer to the real situation. Therefore, the effects of measurement errors in the predictors, and the effects of random predictor variables rather than fixed predictor variables was considered in the model. In the ordinary textbooks, such as Draper and Smith (1966), it is usually assumed that independent (predictor) variables are error free and fixed in the model. However, in psychological and educational research, measurement errors in tests and most other measuring devices are known to exist, and researchers strive to improve their measurements by reducing measurement error or correcting for it. The assumption of error-free predictor variables might reduce conceptual and mathematical labor. However, the results obtained thereby might be quite inappropriate for application to real data. The commonly used measurement theory in psychology assumes that the true score of people have some distribution (Lord and Novick, 1968). Therefore, the assumption of fixed independent (predictor) variables is not realistic when predictor variables are scores on some tests such as aptitude tests. While the error-free fixed predictor variables are attractive and there are numerous textbooks and papers which

use this assumption, it was assumed in this research that the predictor variables are random and have measurement errors.

For predicting a criterion for a future sample, it seems better to express the model as

$$\hat{\underline{y}} = X \underline{b} \qquad (1.1)$$

where $\hat{\underline{y}}$ is a vector of predicted values, X is a matrix of values of predictor variables in the sample where prediction is required, and $\underline{b}$ is a vector of weights. What the model indicates is that the predicted value is a linear combination of predictor values.

Unlike the ordinary linear model, the task here is neither hypothesis testing of some parameters nor estimating the parameters in the model through the experimentally obtained data. The task of prediction is to obtain $\underline{b}$ such that $\hat{\underline{y}}$ is as close to $\underline{y}$ (observed values on a criterion) as possible. The two most popular criteria of closeness are $(\underline{y} - \hat{\underline{y}})'(\underline{y} - \hat{\underline{y}})$ and the correlation between $\underline{y}$ and $\hat{\underline{y}}$. The former criterion is the absolute prediction which is scale dependent, while the latter criterion is relative prediction which does not depend upon scale and is more concerned with the arrangement of people.

Since researchers do not know $\underline{y}$ (otherwise, there is no need to predict), $\underline{b}$ can not be obtained directly. Rather, they are required to use informations in an initial sample to derive weights. A common practice based upon the assumption

that samples are from the same population is to apply the model (1.1) to the initial sample to obtain $\underline{b}$. In the present research, this strategy to obtain $\underline{b}$ was followed with some modification according to the method employed to obtain $\underline{b}$.

CHAPTER II. THEORY OF MEASUREMENT ERRORS AND LEAST SQUARE

Measurement Error

The model used in measurement theory is that the observed score is the sum of true score and error score, i.e.,

$$X = T + E \ , \qquad (2.1)$$

where X is an observed score random variable, T is a true score random variable, and E is an error random variable. It is usually assumed that the expectation of E is zero, the covariance between T and E is zero as well as the covariance among E's. That is,

$$\varepsilon(E)=0,\ \mathrm{Cov}(T, E)=0,\ \mathrm{Cov}(T_1, E_2)=0 \text{ and } \mathrm{Cov}(E_1, E_2)=0 \ ,$$

where two subscripts 1 and 2 denote different measurements (Lord and Novick, 1968; Nunnally, 1967). However, it is preferable to assume that T and E are independent and E's are independent, since zero covariance does not guarantee that there is no relationship between two variables. For example, $y = ax^2 + bx + c$ shows perfect functional relationship between y and x, yet by choosing a, b, and c appropriately, it is possible to obtain zero covariance between x and y. It is, therefore, assumed that the E's are independent of each other and independent of the T's. It should be noted that the assumption of independence implies a zero covariance. In order to avoid the trivial cases, it is also assumed that

variance of X, T, and E are greater than zero and less than infinity.

From the assumptions of independence, it can be easily shown that

$$\mathrm{Var}(X) = \mathrm{Var}(T) + \mathrm{Var}(E) \tag{2.2}$$

and

$$\varepsilon(X) = \varepsilon(T + E) = \varepsilon(T) + \varepsilon(E) = \varepsilon(T). \tag{2.3}$$

The ratio between the true score variance and the observed score variance is called reliability, and is denoted by $\rho_{XX'}$. Thus we have

$$\rho_{XX'} = \mathrm{Var}(T)/\mathrm{Var}(X). \tag{2.4}$$

The probability density function (pdf) of X, $f^*(x)$, can be obtained as follows. Let g(t) and h(e) be the pdf's of T and E, respectively. Then the joint pdf of T and E, f(t,e) is equal to g(t)h(e) because of independence between T and E. Let V be T - E. Then, T = 1/2(X + V) and E = 1/2(X - V). The Jacobian of this transformation is

$$J = \begin{vmatrix} \frac{\partial t}{\partial x} & \frac{\partial t}{\partial v} \\ \frac{\partial e}{\partial x} & \frac{\partial e}{\partial v} \end{vmatrix} = 1/2$$

Therefore, the joint pdf of X and V, $f^*(x,v)$, is equal to $1/2g[1/2(x + v)]h[1/2(x - v)]$ where W and Z are possible ranges of x and v. Then

$$f^*(x) = \int_{v \varepsilon Z} f^*(x,v)\,dv. \tag{2.5}$$

When T and E are normally distributed, $f^*(x)$ is a normal pdf with mean equal to $\varepsilon(T)$ and variance equal to $Var(T) + Var(E)$.

The regression function of X on T, i.e., $\hat{T} = \alpha + \beta X$, which minimizes $\varepsilon(T-\hat{T})^2$ can be obtained as follows. Taking partial derivatives yields

$$\frac{\partial}{\partial\alpha}\varepsilon(T - \alpha - \beta X)^2 = -2\varepsilon(T - \alpha - \beta X) \tag{2.6}$$

$$\frac{\partial}{\partial\beta}\varepsilon(T - \alpha - \beta X)^2 = -2\varepsilon(T - \alpha - \beta X)X. \tag{2.7}$$

Setting (2.6) equal to zero gives

$$\alpha = \varepsilon(T) - \beta\varepsilon(X). \tag{2.8}$$

Setting (2.7) equal to zero and using (2.8) yields

$$\varepsilon(T\cdot X) - \alpha\varepsilon(T) - \beta\varepsilon(X^2)$$

$$= \varepsilon[T\cdot(T + E)] - [\varepsilon(T) - \beta(X)]\varepsilon(T) - \beta\varepsilon(X^2)$$

$$= \varepsilon(T^2) - \varepsilon(T\cdot E) - [\varepsilon(T)]^2 - \beta\{\varepsilon(X^2) - [\varepsilon(X)]^2\}$$

$$= \varepsilon(T^2) - [\varepsilon(T)]^2 - \beta\{\varepsilon(X^2) - [\varepsilon(X)]^2\}$$

$$= Var(T) - \beta Var(X) = 0.$$

Thus β is estimated as

$$\hat{\beta} = Var(T)/Var(X) = \rho_{xx'}.$$

Therefore, $\hat{\alpha} = (1 - \rho_{xx'})\cdot\varepsilon(T)$, and

$$\hat{T} = \rho_{xx'}\cdot X + (1 - \rho_{xx'})\cdot\varepsilon(T). \tag{2.9}$$

Since (2.9) is the equation using the solution to the derivatives it is required to show that the T minimizes the $\varepsilon(T - \tilde{T})^2$. Let $T^* = \alpha^* + \beta^* X$ be any regression function of X.

Then,

$$\varepsilon(T - T^*)^2 = \varepsilon(T - \hat{T} + \hat{T} - T^*)^2$$

$$= \varepsilon(T - \hat{T})^2 + \varepsilon(T - T^*)^2 + 2\varepsilon(T - \hat{T})(\hat{T} - T^*).$$

However,

$$\varepsilon(T - \hat{T})(\hat{T} - T^*)$$

$$= \varepsilon(T - \hat{\alpha} - \hat{\beta}X)[\hat{\alpha} - \alpha^* + (\hat{\beta} - \beta^*)X]$$

$$= (\hat{\alpha} - \alpha^*)\varepsilon(T - \hat{\alpha} - \hat{\beta}X) + (\hat{\beta} - \beta^*)\varepsilon(T - \hat{\alpha} - \hat{\beta}X)X$$

$$= 0 .$$

Therefore,

$$\varepsilon(T - T^*)^2 = \varepsilon(T - \hat{T})^2 + \varepsilon(T - T^*)^2$$

$$\geq \varepsilon(T - \hat{T})^2. \qquad (2.10)$$

Therefore, $\hat{T}$ given by (2.9) minimizes $(T - \tilde{T})^2$. It should be noted that $\hat{T}$ is not an unbiased estimator of $T = t$ unless $\rho_{XX'}$ is equal to unity. The proof is as follows:

$$\varepsilon(\hat{T}/T=t) = \varepsilon[\rho_{XX'} \cdot (t+E) + (1 - \rho_{XX'}) \cdot \varepsilon(T)]$$

$$= \rho_{XX'} t + (1 - \rho_{XX'}) \cdot \varepsilon(T). \qquad (2.11)$$

One of the unbiased estimators of $T=t$ is X since

$$\varepsilon(X/T=t) = \varepsilon(t+E) = t. \qquad (2.12)$$

However, X has larger mean square error than $\hat{T}$, and $\hat{T}$ has been used as an estimator of T in the psychological and educational fields.

The Least Square Method and a Linear Model

A linear model is defined as

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \qquad (2.13)$$

$i=1, 2, \ldots, n$

or using matrix notation,

$$\underline{y} = X\underline{\beta} + \underline{e} \qquad (2.14)$$

where $\underline{y}=(y_i)$ is an nxl dependent variable vector (observations), $\underline{\beta}=(\beta_j)$ is a pxl parameter vector, X is an nxp matrix of independent variables with rank p and $\underline{e}=(e_i)$ is an nxl error vector. It is usually assumed through random sampling that

$$\varepsilon(\underline{e}) = \underline{0} \qquad (2.15)$$

and

$$\mathrm{Var}(\underline{e}) = \sigma^2 I. \qquad (2.16)$$

The model is linear in the unknown parameter vector ($\underline{\beta}$).

The well-known parts of the theories of linear models are based upon the case where the X matrix is an error-free fixed matrix. However, Hocking (1976) made a comment on the assumption of a fixed X matrix as follows:

> The input x_{ij} are frequently taken to be specific design variables, but in many cases it is more appropriate to consider them as random variables and assume a joint distribution on y and x, say multivariate normal.

The least square estimation in the linear model under the assumption of fixed X is defined as finding $\underline{b}$ such that $(\underline{y} - X\underline{b})'(\underline{y} - X\underline{b})$ is minimum. Under the assumptions of (2.15) and (2.16), least square estimation gives the following properties,

$$\underline{b} = (X'X)^{-1}X'\underline{y} \tag{2.17}$$

$$\varepsilon(\underline{b}) = \underline{\beta} \tag{2.18}$$

$$\mathrm{Var}(\underline{b}) = \sigma^2(X'X)^{-1}. \tag{2.19}$$

A predicted value of y for a future observation, given that predictor variables have value $\underline{x}$, is given by

$$\hat{y} = \underline{x}'\underline{b} \tag{2.20}$$

and

$$\mathrm{Var}(\hat{y}) = \sigma^2(1 + \underline{x}'(X'X)^{-1}\underline{x}). \tag{2.21}$$

These equations given above are well-known among researchers in psychology and education, and they are quite often used without consideration of the nature of predictor variables, namely, the X matrix.

When the X matrix is assumed to be random, the equations (2.17) through (2.21) can be considered as the conditional equation on the sample. Let $\underline{x}_i'$ be an i^{th} row of the matrix X and a random vector. When $\underline{x}_i'$ and e_i are independently and normally distributed, y_i and $\underline{x}_i'$ have a multivariate normal distribution. Let $\underline{z}_i$ be $(y_i, \underline{x}_i')'$ from a multivariate normal distribution with mean $(\mu_y, \underline{\mu}_x')'$ and variance $\begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then

the least square estimation on the deviate scores, defined as $(\underline{z}_i - \bar{\underline{z}})$ where $\bar{\underline{z}}$ is a sample mean, is the same as the maximum likelihood estimation (Anderson, 1958; Graybill, 1976; Narula, 1974; Sampson, 1974) and given by

$$\underline{b} = S_{22}^{-1} s_{21} \tag{2.22}$$

where $S = \begin{pmatrix} s_{11} & s_{12} \\ s_{21} & S_{22} \end{pmatrix}$ is a sample sum of squares and cross products matrix defined as $\sum_{i=1}^{n} (\underline{z}_i - \bar{\underline{z}})(\underline{z}_i - \bar{\underline{z}})'$. The sample variance matrix also can be used to derive $\underline{b}$ instead of S because of the cancellation of divisors. The regression equation is given by

$$\hat{y} = \bar{y} + (\underline{x} - \bar{\underline{x}})'\underline{b} \tag{2.23}$$

where $(\bar{y}, \bar{\underline{x}}')' = \bar{\underline{z}}$. Sampson (1974) gave the following properties obtained from multivariate normality.

$$\varepsilon(\underline{b}) = \underline{\beta} = \Sigma_{22}^{-1} \Sigma_{21}. \tag{2.24}$$

$$\mathrm{Var}(\underline{b}/S_{22}) = S_{22}\sigma^2 \quad \text{where } \sigma^2 = \sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{2.25}$$

$$\mathrm{Var}(\underline{b}) = \Sigma_{22}^{-1}\sigma^2/(n-p-1). \tag{2.26}$$

It should be noted that $\varepsilon(S_{22}) = (n-p-1)^{-1}\Sigma_{22}^{-1}$ because S_{22} has a Wishart distribution.

Narula (1974) showed that the predicted mean square error conditional on $\underline{x}$ is

$$\varepsilon[(y-\hat{y})^2/\underline{x}] = \sigma^2\left(1+\frac{1}{n}\right) + \frac{\sigma^2[(\underline{x}-\underline{\mu}_x)'\Sigma_{22}^{-1}(\underline{x}-\underline{\mu}_x) + \frac{p}{n}]}{n-p-2}. \tag{2.27}$$

Kerridge (1967) and Stein (1961) showed that the expectation of the mean square error is

$$\varepsilon(y-\hat{y})^2 = \sigma^2(1 + \frac{1}{n})(n-2)/(n-p-2) \quad . \tag{2.28}$$

Kerridge (1967) also showed that the covariance between the error of prediction of y_{n+1} and y_{n+2} is

$$\varepsilon(y_{n+1} - \hat{y}_{n+2})(y_{n+2} - \hat{y}_{n+2}) = \frac{\sigma^2(n-2)}{(n-p-2)n} \quad . \tag{2.29}$$

As equations (2.22) through (2.29) show, the estimation of $\underline{\beta}$ is the same for fixed X or random X, yet the other properties are quite different, such as Var($\underline{b}$) and the mean square error of a prediction.

When the sample size, n, increases, the sample variance matrix becomes a better estimator of the parameter and $\underline{b}$ approaches $\underline{\beta}$. This can be seen in equation (2.26) where Var($\underline{b}$) approaches zero as n increases. Also when n increases, the predicted mean square error conditional on $\underline{x}$, given by (2.27), and the mean square error, given by (2.28) approaches σ^2, while the covariance of errors of prediction y_{n+1} and y_{n+2} approaches zero. Therefore, the least square method can be a good method when the sample size is large. However, it can be seen from the equations (2.26) through (2.29) that the ratio between n and p is important to determine goodness of prediction when n is small.

The above discussions and properties are based upon the assumption of multivariate normality. Graybill (1976) derived the maximum likelihood estimator of $\underline{\beta}$ when $\underline{x}$ has the unknown distribution and e_i has the normal distribution. It was also given by the equation (2.22) which can be obtained through the least square method. The case where $\underline{x}$ and e_i have an unknown distribution was also discussed and some suggestions were given by Graybill (1976).

CHAPTER III. ALTERNATIVE METHODS

Ridge Regression

Since Hoerl and Kennard (1970a, 1970b) proposed ridge regression and its application, numerous research papers have appeared, addressing theories, criticisms, generalization and modification of ridge ideas. The main concern of ridge regression is that, when the obtained data of independent variables are correlated, the least square solutions are sometimes unsatisfactory or do not make sense. For example, some of the b_i's might have signs contrary to those which are expected from the theory. Ridge regression was proposed to compensate for the ill-conditioned variance matrix resulting from the correlated independent variables.

The model (2.14) with assumptions (2.15) and (2.16) is used to develop ridge regression, that is,

$\underline{y} = X\underline{\beta} + \underline{e}$ with the rank of X equal to p,

$\varepsilon(\underline{e}) = \underline{0}$, and $\mathrm{Var}(\underline{e}) = \sigma^2 I$.

In the theory, X and $\underline{y}$ are transformed to standard scores with mean zero and variance one in the sample. Since the theory was developed for the case where the X matrix was fixed, the following discussion assumes the X matrix is fixed. In the case of the random X, the following discussion can be taken as conditional on the initial sample. However,

equations (2.22) through (2.26) provide easy translation from fixed X to random X.

As discussed in the previous section, the least square estimator of $\underline{\beta}$ is given by

$$\underline{b} = (X'X)^{-1}X'\underline{y}.$$

It should be noted that X and $\underline{y}$ are standardized and X'X is a form of a sample correlation matrix among predictor variables and X'$\underline{y}$ is a form of sample correlations between predictor variables and a criterion variable (validity coefficients). Let the distance between $\underline{b}$ and $\underline{\beta}$ be L_1 and

$$L_1^2 = (\underline{b} - \underline{\beta})'(\underline{b} - \underline{\beta}). \tag{3.1}$$

Then,

$$\varepsilon(L_1^2) = \sigma^2 tr(X'X)^{-1} \tag{3.2}$$

or equivalently

$$\varepsilon(\underline{b}'\underline{b}) = \underline{\beta}'\underline{\beta} + \sigma^2 tr(X'X)^{-1} . \tag{3.3}$$

When the error $\underline{e}$ is normally distributed,

$$Var(L_1^2) = 2\sigma^4 tr(X'X)^{-2} . \tag{3.4}$$

When the eigenvalues of X'X are denoted by

$$\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p = \lambda_{min} > 0,$$

$$\text{then, } \varepsilon(L_1^2) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \tag{3.5}$$

and

$$\mathrm{Var}(L_1^2) = 2\sigma^4 \sum_{i=1}^{p} \left(\frac{1}{\lambda_i}\right)^4. \tag{3.6}$$

Equation (3.5) shows that when the smallest eigenvalue, λ_{min}, is very small relative to the other eigenvalues the distance between $\underline{b}$ and $\underline{\beta}$ can be very large.

In order to deal with the small value of λ_{min} and reduce the large $\varepsilon(L_1^2)$, it was first proposed by Hoerl (1962) to use an estimator $\underline{b}^*$ given by

$$\underline{b}^* = (X'X + kI)^{-1}X'\underline{y}, \; k \geq 0. \tag{3.7}$$

The relationship between a ridge estimator to a least square estimator is given by the alternative forms,

$$\underline{b}^* = [I + k(X'X)^{-1}]^{-1}\underline{b} \tag{3.8}$$

and

$$\underline{b}^* = (X'X + kI)^{-1}X'X\underline{b}. \tag{3.9}$$

The characteristic of $\underline{b}^*$ is that when the squared length of the estimator of $\underline{\beta}$ is fixed, $\underline{b}^*$ gives the minimum sum of squares of the residuals. However, it can be easily obtained from (3.8) or (3.9) that $\underline{b}^*$ is a biased estimator of $\underline{\beta}$ for k not equal to zero. Hoerl and Kennard (1970a) showed that there exists k such that

$$\varepsilon[L_1^2(k)] < \varepsilon[L_1^2(0)] = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$$

where $\varepsilon[L_1^2(k)] = \varepsilon(\underline{b}^* - \underline{\beta})'(\underline{b}^* - \underline{\beta})$ with $\underline{b}^*$ from (3.7). Therefore, $\underline{b}^*$ is a biased estimator with a shorter distance from the parameter $\underline{\beta}$ on the average than the least square estimator. Hoerl and Kennard (1970a) further developed ridge regression into generalized ridge regression. Let P be the orthogonal matrix such that X'X=PSP' where S is a diagonal matrix with eigenvalues (λ_i's) of X'X. Let

$$X^* = XP, \ \underline{\alpha} = P'\underline{\beta} \text{ and } \underline{y} = X^*\underline{\alpha} + \underline{e}. \tag{3.10}$$

Then the generalized ridge estimation is defined as

$$\underline{\alpha} = [(X^*)'(X^*) + K]^{-1}(X^*)'\underline{y} \tag{3.11}$$

where K is a diagonal matrix with k_{ii} where i varies from 1 through p. The optimal value of k_{ii} is σ^2/λ_i^2, but the numerator is the parameter which is unknown. Hemmerle (1975) obtained the explicit solution for k_{ii}'s. Guilkey and Murphy (1975) modified the generalized ridge estimator such that only the elements with small eigenvalues are manipulated by k_{ii}'s. The diagonal λ_i's which are less than $10^{-c}\lambda_{max}$ are manipulated by adding k_{ii}'s and the rest of k_{jj}'s are set to be zero.

Marquardt (1970) discussed the relationship between ridge regression and the generalized inverse with continuous rank $r \leq p$. The generalized inverse A^+ of X'X with rank r is defined as

$$A^+ = \sum_{i=1}^{a} \frac{1}{\lambda_i} \underline{p}_i\underline{p}_i' + (r-a)\, \underline{p}_{a+1}\underline{p}_{a+1}'/\lambda_{a+1} \tag{3.12}$$

where $\underline{p}_i$ is an eigenvector corresponding to the eigenvalue λ_i and a is the largest integer less than or equal to r. When r is an integer the regression is called components regression or orthogonalized regression. Marquardt (1970) summarized the relationship as:

> The Ridge and Generalized Inverse estimators share many properties. Both are superior to least square for all conditioned problems. For both classes of estimators the degree of bias (choice of k or r) can be bracketed within a reasonable range in any given instance and practical results can be obtained. The generalized inverse solution is especially relevant for precisely zero eigenvalues. The ridge solution is computationally simpler and it seems better suited to coping with very small, but nonzero eigenvalues.

Marquardt and Snee (1975) showed in their simulation study that the ridge regression and the generalized inverse methods are superior to the least square method, especially when the independent variables are highly correlated. Hawkins (1975) showed that ridge regression is a weighted sum of the eigenvectors and he discussed the extension to ridge regression.

The major problem of ridge regression, however, is how to decide the value of k in (X'X+kI). That is, what criterion is used to determine the value of k? Hoerl and Kennard

(1970b) suggested the following steps:

1) Compute $\underline{b}^* = (X'X + kI)^{-1}X'\underline{y}$ for certain values of k.

2) Plot b_i^*'s against the values of k (Ridge trace).

3) Find a value of k where the system stabilizes.

Marquardt and Snee (1975) commented about the choice of the value of k as follows:

> Many statisticians have expressed concern about the selection of k. It is the authors' experience that this is not a problem in practice. As will be pointed out later in the examples. The plot of prediction standard deviation of new data versus k usually has a flat minimum; hence, there is a range of k-values which give equivalent results from a practical point of view.

Obenchain (1975), however, showed that different criteria result in a wide range of k values for the same set of data. He showed that the 10-factor data by Gorman and Toman (1966) can have k values ranging from 0.008 to 0.54 depending upon the criterion used. Visual inspection of the ridge trace by Hoerl and Kennard (1970b) resulted in k value between .2 and .3 which was the second highest value of k in Table 3 by Obenchain (1975). Obenchain recognized two kinds of criteria to determine the value of k. One criterion judges goodness of $X\underline{b}^*$ as a predictor of $\underline{y}$ while the other criterion attempts to evaluate $\underline{b}^*$ with a minimum of reference to data. The former criterion is called the prediction oriented criterion

and tends to result in smaller values of k, while the latter criterion is called the control oriented criterion and tends to result in the larger values of k. It should be noted that the word "prediction" used by Obenchain does not refer to prediction in the future sample with random X, but it refers to the prediction with fixed X. Klingler (1975) used the visual inspection method in his simulation study of prediction in the second sample with the fixed X matrix and showed that ridge regression is generally better than least square, especially when the sample size is small.

When the X matrix is random instead of fixed, the equation (2.25) or (2.26) can be used according to the definition of prediction, i.e., conditional on the initial sample or not, in order to obtain L_1^2 in (3.2). Therefore, properties of $\underline{b}^*$ based upon the fixed X matrix can be extended to the case where the X matrix is random. In particular, ridge regression will still result in a biased but closer estimator of $\underline{\beta}$ in the case of random X.

Equal Weighting

Weighting systems of linear functions of correlated variables were first explored by Wilks (1938). One of his findings was that under certain reasonable conditions the mean value of the correlation between the two linear functions of p variables differs from unity, by terms of order 1/p,

and the variance of the correlation is of order $1/p^2$. Gulliksen (1950) obtained more general cases of the weighting system. According to Perloff (1951), general characteristics of the weighting system are:

1) The effect of a (psychological) test on a composite is based not on its mean and its length, but rather on the total ranges of scores;
2) Weighting will be ineffectual when p, $\bar{r}$, and the mean weights are relatively high, while the variance of the weights is low.

Therefore, it can be expected that when the predictor variables are highly correlated, any weighting system can perform reasonably well.

The early empirical studies by Lawshe and Schucker (1959), Perloff (1951), and Wesman and Bennett (1959) showed that the equal weighting system after adjusting the sign performed as well as the least square weighting for prediction of a criterion variable in the second sample. It should be noted that two kinds of equal weighting systems were used in these early studies. One is the sum of raw scores (Lawshe and Schucker, 1959; Wesman and Bennett, 1959) and the other is the sum of the standardized score (Perloff, 1951). In the latter case, weights on the raw scores are the inverse of their standard deviations. Perloff (1951) concluded that, with a small sample size, standardized-score unit weights

(sum of standardized scores) yielded better prediction than the weights obtained by the least square method in terms of the correlation between the predicted values and observed values on the criterion variables (relative prediction). As the sample size increases, the least square method starts to perform as well as or better than other methods.

Schmidt (1971) showed that unit weighting, that is all weights equal to unity, might be superior to the weights obtained by the least square method. Schmidt's simulation study showed that when the sample size was equal to 25, unit weighting performed better where the number of predictors was 2, 4, 6, 8 and 10 with and without suppressor variables. When there was no suppressor variable, equal weights performed better for sample size less than or equal to 150 under certain conditions. In general, the data suggested that as the number of predictor variables increases, it requires a greater sample size for the least square method to perform better than equal weighting. Dawes and Corrigan (1974) presented four studies where unit weighting did extremely well in predicting the criterion values. One of their findings was that the predictor variables had conditionally monotone relationship to the criterion variable or may easily be scaled to have such a relationship. Therefore, the additive model is a fairly good approximation of other monotonic functions. Yet, since the additive model is an

approximation, it might be possible that the equal weighting is more robust against deviation from the model than differentially weighting each variables.

Einhorn and Hogarth (1975) investigated differences between equal weighting and the least square method. In particular, they investigated the models

$$y_i = \alpha + \sum_{j=1}^{p} \beta_j x_{ij} + e_i \quad i=1, 2,..., n \tag{3.13}$$

and

$$y_i = \delta + \gamma \sum_{j=1}^{p} x_{ij} + u_i \tag{3.14}$$

They focused attention on the mean square error rather than the sum of squared residuals. Since the model (3.13) has degrees-of-freedom (df) associated with error of n-p-1 while the model (3.14) has df associated with error of n-2, it can be easily shown that the model (3.14) can predict better than the model (3.13) does under certain conditions. The relative performance of the two models is related to variances and covariances, the sample size and the number of predictor variables.

Green (1977) looked at the loss caused by deviation from the least square weights, such as rounding effect. He found that the loss was, most of the time, small. Wainer (1976, 1978) and Wainer and Thissen (1976) emphasized that equal weighting can perform as well as other methods, if not better, and it is a very robust procedure.

CHAPTER IV. DEVELOPMENT OF THE PROBLEMS

The major problems of the least square method applied to prediction of psychological criteria are multicollinearity among the predictors, the criterion of fit and measurement errors in the predictor variables. The least square method, which minimizes $(\underline{y} - X\underline{b})'(\underline{y} - X\underline{b})$, depends heavily upon the condition of the X'X matrix. When the predictor variables are highly correlated, the X'X matrix might approach a singular matrix. Yet, it is hard to detect near singularity (ill-condition) from data or the X'X matrix. When X is fixed or random but conditional on the initial sample, equation (3.5) shows that the squared distance between $\underline{b}$ and $\underline{\beta}$ is inversely related to the eigenvalues of X'X. When the predictor variables are orthogonal and scores are standardized, X'X is close to a diagonal matrix, i.e., I, and the eigenvalues are close to one. As X'X approaches a singular matrix, the ratio between the largest eigenvalue and the smallest eigenvalue increases, and the average L_1^2 in (3.5) also increases. It is not rare for psychological and educational researchers to use moderately or highly correlated predictors. Therefore, $\underline{b}$ obtained by the least square method from the initial sample might be quite useless for predicting the criterion variable in the second sample, because of unstable weights. For an unconditional case, the same argument

follows where L_1^2 is inversely related to the eigenvalues of Σ_{22}.

Stein (1960) used the multivariate normal distribution to investigate the nature of the maximum likelihood (ML) estimator of the weights. It should be noted that the ML estimator is the same as the least square estimator as discussed in Chapter II. Let $\underline{z}_1, \ldots, \underline{z}_n$ be independently normally distributed (1+p) dimensional random vectors defined in Chapter II. Then,

$$\varepsilon(y_i/\underline{x}_i) = \underline{x}_i' \underline{\beta} + \alpha \tag{4.1}$$

where $\underline{\beta} = \Sigma_{22}^{-1}\Sigma_{21}$ and $\alpha = \mu_y - \underline{\mu}_x'\underline{\beta}$.

Let $\hat{\underline{\beta}}$ and $\hat{\alpha}$ be ML estimators of $\underline{\beta}$ and α given by (2.22) and (2.23). Stein defined the measure of the error, the risk, as

$$L = \frac{[(\hat{\alpha}-\alpha) + (\hat{\underline{\beta}}-\underline{\beta})'\underline{\mu}_x]^2 + (\hat{\underline{\beta}}-\underline{\beta})'\Sigma_{22}(\hat{\underline{\beta}}-\underline{\beta})}{\sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}} \tag{4.2}$$

The expected risk of the ML estimators is given by

$$\varepsilon(L) = \begin{cases} \dfrac{n(p+1) - 2}{n(n-p-2)} & \text{if } n \geq p+3 \\ \infty & \text{if } n \leq p+1 . \end{cases} \tag{4.3}$$

When μ_y and $\underline{\mu}_x$ are known, the expected risk is given by

$$\varepsilon(L) = \begin{cases} \dfrac{p}{n-p-1} & \text{if } n \geq p+2 \\ \infty & \text{if } n \leq p+1 . \end{cases} \tag{4.4}$$

Stein (1960) found that for the known means, $p = 3$, $n \geq p+2$ the ML estimator $\hat{\underline{\beta}}$ is not an admissible estimator and for

sufficiently small a and d (with both a, d > 0) the estimator

$$\underline{b} = [1 - \frac{d(1 - R^2)}{a(1 - R^2) + R^2}]\hat{\underline{\beta}} \ . \tag{4.5}$$

has everywhere smaller risk than $\hat{\underline{\beta}}$ where R^2 is the squared simple multiple correlation coefficient. Stein tentatively recommended use of the estimator

$$\underline{b} = \max\ [(1 - \frac{p - 2}{n-p+2}\frac{1 - R^2}{R^2}),\ 0]\hat{\underline{\beta}}. \tag{4.6}$$

The recommended $\underline{b}$ is uniformly reduced in absolute value toward the origin. The approach taken by Stein is very similar to ridge regression in the sense that when a criterion of fit is altered, least square is not necessarily the best method.

Narula (1974) showed that a smaller number of predictor variables and the reduced least square weights have smaller residual sum of squares than the full model least square weights when the weights obtained in the initial sample were applied to the second sample to obtain the predicted values. Narula gives the prediction equation of the reduced least square weights as follows:

$$\hat{y}_i = \bar{y} + \lambda_{\underline{x}_i}(\underline{x}_i - \bar{x})'\underline{b} \tag{4.7}$$

where $\underline{b}$ is the least square estimator, $\underline{x}_i$ is the vector of the values of predictor variables, $\bar{y}$ and $\bar{x}$ are means obtained from the initial sample, and $0 \leq \lambda_{\underline{x}_i} \leq 1$ is a function of $\underline{x}_i$, $\underline{b}$ and the sample variance matrix.

The effects of measurement error on the least square estimates have not been studied much by psychologists. Wolins (1967) recognized two sources of variability in weights. One is the sampling error accounted for by the usual procedure and the other is the measurement error. Economists have been concerned with the effects of errors in the variables and studied the error-in-variables model (Johnson, 1963; Malinvaud, 1966). Schneeweiss (1976) derived a consistent estimator for the case of known variance and covariances of error variables. He also obtained the asymptotic distribution of the estimator. Warren et al., (1974) used an error-in-variables model to analyze the managerial role performance. However, they warn against the use of the estimator of $\underline{\beta}$ derived from the error-in-variables model to predict the future observation of the criterion variable. The purpose of the error-in-variables model is for estimation and understanding of structure and not prediction. This can be easily seen in the following example. Let y and t_j's, j=1, 2, and 3, be random variables without error and e_j's, j=0, 1, 2, 3, be errors for y and t_j's, which are independent of each other and y and t_j's with mean 0. Let the true model be

$$y = t_1 + t_2 + t_3 + d \tag{4.8}$$

where d is a deviation or error and its expectation is zero. Since observed values of predictor variables are $(t_j + e_j)$'s,

the true weights in the model (4.8), i.e., unities, would not help to predict y in the future sample unless variances of e_j's are equal. If e_j's have different variances, the prediction weights have to reflect these different variances and the reliable variables have to be weighted more than unreliable variables under the model (4.8).

Let $(y, \underline{t}')$ be jointly distributed with mean $\underline{0}$ and variance $\begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, where $\underline{t}' = (t_1, \ldots, t_p)$ is a true score vector of the predictor variables. Let $\underline{x}$ be $\underline{t} + \underline{e}$ where $\underline{e}$ is a measurement error vector with the assumptions in Chapter II. It should be noted that normality is not assumed here. Let the true model be

$$y = \alpha + \underline{t}'\underline{\beta} + d \qquad (4.9)$$

where d is a deviation with mean 0 and $\mathrm{Var}(d) = \sigma^2$. Let the prediction model be

$$\hat{y} = \eta + \underline{x}'\underline{\gamma}. \qquad (4.10)$$

Then,

$$\begin{aligned} \varepsilon[(\hat{y}-y)^2/\underline{t}] &= \varepsilon[(\eta + \underline{x}'\underline{\gamma} - \alpha - \underline{t}'\underline{\beta} - d)^2/\underline{t}] \\ &= \varepsilon[\{(\eta - \alpha) + (\underline{t} + \underline{e})'\underline{\gamma} - \underline{t}'\underline{\beta} - d\}^2/\underline{t}] \\ &= \varepsilon[\{(\eta - \alpha) + \underline{t}'(\underline{\gamma} - \underline{\beta}) + \underline{e}'\underline{\gamma} - d\}^2/\underline{t}] \\ &= (\eta - \alpha)^2 + [\underline{t}'(\underline{\gamma} - \underline{\beta})]^2 + \underline{\gamma}'G\underline{\gamma} \\ &\quad + 2(\eta - \alpha)\,\underline{t}'\,(\underline{\gamma} - \underline{\beta}) + \sigma^2, \end{aligned} \qquad (4.11)$$

where $G = \mathrm{Var}(\underline{e}) = \mathrm{diag}(g_{ii})$ and $g_{ii} > 0$. By unconditioning with respect to $\underline{t}$,

$$\varepsilon(\hat{y} - y)^2 = (\eta - \alpha)^2 + (\underline{\gamma} - \underline{\beta})'\Sigma_{22}(\underline{\gamma} - \underline{\beta})$$

$$+ \underline{\gamma}'G\underline{\gamma} + \sigma^2. \tag{4.12}$$

Taking the derivative with respect to α and $\underline{\beta}$ and setting them equal to zero gives

$$\eta = \alpha \tag{4.13}$$

and

$$\underline{\gamma} = (G + \Sigma_{22})^{-1}\Sigma_{22}\underline{\beta}. \tag{4.14}$$

It can be easily shown that η and $\underline{\gamma}$ minimizes (4.12). When the multivariate normality of d, $\underline{t}$, and $\underline{e}$ is assumed, $(y, \underline{x}')$ is also distributed with mean $\underline{0}$ and variance $\begin{pmatrix} \sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} + G \end{pmatrix}$. Then,

$$\underline{\beta} = \Sigma_{22}^{-1}\Sigma_{21}$$

and

$$\underline{\gamma} = (G + \Sigma_{22})^{-1}\Sigma_{21} \cdot \tag{4.15}$$

Presence of measurement errors in prediction variables affects $\mathrm{Var}(\underline{b}/S_{22})$ and $\mathrm{Var}(\underline{b})$ given by (2.25) and (2.26) in two ways. Firstly, it reduces multicollinearity among predictor variables because the diagonals are inflated by G. The ratio $\lambda_{max}/\lambda_{min}$ will be reduced, which supposedly stabilizes $\underline{b}$. Secondly, measurement errors increase the variance of lack of fit to σ^2. When there is no measurement error, $\sigma^2 = \sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$, but it changes to $\sigma_{11} - \Sigma_{12}(\Sigma_{22} + G)^{-1}\Sigma_{21}$ when measurement error exists. The

proof of increase in σ^2 can be easily obtained from the following result by Rao (1965).

Let A and D be nonsingular matrices of orders m and n and B be mxn matrix. Then

$$(A+BDB')^{-1} = A^{-1} - A^{-1}B(B'A^{-1}B+D^{-1})^{-1}B'A^{-1}.$$

By setting B = I, n = m =p, A = Σ_{22} and D = G, it is readily seen that

$$\Sigma_{12}(\Sigma_{22} + G)^{-1}\Sigma_{21}$$

$$= \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \Sigma_{12}\Sigma_{22}^{-1}(\Sigma_{22}^{-1} + G^{-1})^{-1}\Sigma_{22}^{-1}\Sigma_{21}.$$

Since Σ_{22} and G are positive definite and $\Sigma_{12} = \Sigma_{21}'$, the second term of the right hand side is greater than zero. Therefore,

$$\Sigma_{12}(\Sigma_{22} + G)^{-1}\Sigma_{21} < \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

$$\sigma_{11} - \Sigma_{12}(\Sigma_{22} + G)^{-1}\Sigma_{21} > \sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

As equations (3.7) and (4.15) show, there exists a strong structural similarity between the case with measurement error and ridge regression. The equation (4.15) suggests that the diagonals of less reliable predictor variables in the variance matrix are inflated. Since ridge regression (3.7) inflates the diagonals to reduce the ill-condition of the X'X matrix, the diagonal matrix G in (4.15) might be able to replace I in (3.7) to reflect the precision of the predictor variables. Therefore, it is proposed to

modify the ridge regression as

$$\underline{b}^{*} = (X'X + kG)^{-1}X'\underline{y} \tag{4.16}$$

where G is the diagonal matrix of $(1 - \rho_{xx'})$'s, that is, one minus the reliability of the predictor variable, and X and $\underline{y}$ are standardized scores. When the true reliabilities are not available, estimates of them are used. Therefore, modified ridge regression would be, in practice,

$$\underline{b}^{**} = (X'X + k\hat{G})^{-1}X\underline{y}. \tag{4.17}$$

In the first study, least square, modified ridge regression and ordinary ridge regression were compared. Also, the effect of the different values of k was investigated further since the study of k values by Marquardt and Snee (1975) was based upon only one case.

In the second study, the least square method, the equal weighting method and ridge regression were investigated. Ridge regression was the ordinary ridge regression given by (3.7), based upon the result of the first study. In the second study, the effect of restriction on the sampling was also investigated as one of the possible violations of the assumptions in order to find the robustness of the methods. In psychological research and other applied fields, restriction on the initial sample is one of the most common problems. For example prediction of the first year freshman GPA from the predictors might face a problem of dropout of the low

ability students in the sample. The restricted sample alters the distribution of true scores while the distribution of measurement error is assumed to stay as before.

CHAPTER V. STUDY 1

Method

Population Parameters

The number of predictor variables in the study was set to be 4 and 8. In order to obtain the variance matrix for the predictor variables, the last 8 variables of Table 2 in Ayers (1971) were used. The diagonal entries were reduced, relative to the nondiagonal entries, by setting them equal to the maximum correlation in the column. The smallest eigenvalue of the matrix was manipulated to obtain a nonsingular matrix. Then the matrix was rescaled with ones in the diagonals. The same procedure was followed to obtain the 4x4 variance matrix using the last four variables of Table 2 in Ayers (1971). These variance matrices are given in Table 1 and Table 2.

Table 1. Variance matrix of 4 predictors without error

	Var1	Var2	Var3	Var4
Var1	1.00	0.85	0.68	0.59
Var2		1.00	0.94	0.89
Var3			1.00	0.87
Var4				1.00

Table 2. Variance matrix of 8 predictors without error

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8
Var1	1.00	0.57	0.68	0.66	0.37	0.66	0.67	0.86
Var2		1.00	0.72	0.69	0.44	0.63	0.86	0.68
Var3			1.00	0.72	0.24	0.71	0.70	0.75
Var4				1.00	0.36	0.79	0.79	0.72
Var5					1.00	0.70	0.74	0.64
Var6						1.00	0.86	0.80
Var7							1.00	0.81
Var8								1.00

The eigenvalues of the 4x4 matrix were 3.42, 0.45, 0.12 and 0.0084, and those of the 8x8 matrix were 5.76, 0.92, 0.53 0.36 0.25, 0.102, 0.05 and 0.0097. Two vectors of variance of error variables are given in Table 3 and Table 4, with the reliabilities, the largest eigenvalue (λ_{max}), the smallest eigenvalue (λ_{min}) of the population variance matrix with diagonals increased by error variances, and the correlation ($R_{y\hat{y}}$) between observed variable (y) and predicted value ($\hat{y}$) using weights derived from (4.15). The validity coefficients, i.e., the correlations between predictors and the criterion, were taken from the first row of the Table 2 in Ayers (1971), which were

(0.46 0.54 0.32 0.25 0.23 0.35 0.47 0.53)

Table 3. Error variances for 4 predictors

	Group D41		Group D42	
	Variance	Reliability	Variance	Reliability
Var1	0.23	0.813	0.13	0.885
Var2	0.10	0.909	0.05	0.952
Var3	0.10	0.909	0.05	0.952
Var4	0.15	0.870	0.10	0.909
λ_{max}	3.5624		3.5017	
λ_{min}	0.1222		0.0685	
$R_{y\hat{y}}$	0.5455		0.5871	

Table 4. Error variances for 8 predictors

	Group D81		Group D82	
	Variance	Reliability	Variance	Reliability
Var1	0.24	0.806	0.14	0.877
Var2	0.15	0.870	0.05	0.953
Var3	0.37	0.730	0.27	0.787
Var4	0.25	0.800	0.15	0.870
Var5	0.31	0.763	0.21	0.826
Var6	0.13	0.885	0.03	0.971
Var7	o.15	0.870	0.05	0.952
Var8	0.24	0.806	0.14	0.877
λ_{max}	5.9861		5.8861	
λ_{min}	0.1871		0.0872	
$R_{y\hat{y}}$	0.6142		0.6571	

for 8-predictor cases, and the last four of them were the validity coefficients for 4-predictor cases. The variance of the criterion variable was set to be 1.00. Means of variables were all set to zero.

Factors Manipulated

Five factors were manipulated in the first study. Factor one was the number of predictor variables and factor two was the degree of measurement errors in the predictor variables as given by Table 1 through Table 4. Factor three was the sample size, n. Two levels of sample size, n=25 and n=75 were chosen. Factor four was the method to obtain prediction weights. The least square method and ridge regressions given by (3.7) and (4.17) were compared. Ridge regression based upon (3.7) will be referred to as ordinary ridge regression whereas the second one based upon (4.17) will be referred to as modified ridge regression. Factor five was the value of k in ridge regressions. Values of k were set to 0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.2, 0.3, 0.4, 0.6, 0.8, and 1.0. Diagonal elements of $\hat{G}$ in (4.17) were smaller than one. Therefore, values of k given above were divided by one minus the mean of the reliabilities of the predictor variables for modified ridge regression.

Procedure

Twelve sets of 25 independent vectors and 75 independent vectors were generated from the multivariate normal distribution with mean $\underline{0}$ and variance matrix augmented by the criterion variable. The diagonal elements of the variance matrix for the predictor variables were sums of one and error variances. For each set of data, obtained scores were standardized with means equal to zero and variances equal to one before (3.7) and (4.17) were applied with different k values. For least square, k was set to zero. $\hat{G}$ in (4.17) was generated from p χ^2-distributions with n-1 degree of freedom. The value of the i^{th} element in $\hat{G}$ was computed from

$$(1 - \rho_{x_i x_i'})\chi^2_{n-1}/(n-1). \qquad (5.1)$$

Weights obtained were then converted back to original scales using sample means and standard deviations.

Two types of criteria for fit of prediction were calculated for each set of data. One was the correlation between the predicted value $\hat{y}$ and the observed value y, denoted by $\rho_{y\hat{y}/\underline{b}}$, which corresponds to relative prediction. $\rho_{y\hat{y}/\underline{b}}$ was given by

$$\rho_{y\hat{y}/\underline{b}} = \frac{\mathrm{Cov}(y, \hat{y})}{\sigma_y \sigma_{\hat{y}}}$$

$$= \frac{\sum_{j=1}^{p} b_j \sigma_{yx_j}}{\sqrt{\sum_{i=1}^{p} b_i \sigma^2_{x_i} + 2 \sum_{a=1}^{} \sum_{\substack{c=1 \\ c \neq a}}^{} b_a b_c \sigma_{x_a x_c}}} \qquad (5.2)$$

where x_i's are predictor variables, p is the number of predictor variables and b_i's are the obtained weights. Another criterion was the expected value of squared prediction residual given by

$$\varepsilon(y - \hat{y})^2/\underline{b} = \sigma_y^2 + b_0^2 + \sum_{i=1}^{p} b_i \sigma^2_{x_i} + 2 \sum_{a=1}^{} \sum_{\substack{c=1 \\ c \neq a}}^{} b_a b_c \sigma_{x_a x_c}$$

$$- 2 \sum_{j=1}^{p} b_j \sigma_{yx_j}, \qquad (5.3)$$

which corresponds to absolute prediction.

There were four different variance matrices which generated 12 sets of 25 independent vectors and 12 sets of 75 independent vectors. These four groups of data were denoted as group D41, D42, D81, and D82 as shown in Table 3 and Table 4.

Results and Discussion

For each set of data, the best values of two criteria were selected over the different values of k for ordinary and modified ridge regressions. The best value of the

correlation between y and $\hat{y}$ was the highest correlation over correlations among different values of k, whereas the best value of the expected value of squared prediction residual was the smallest one. The numbers of sets where ordinary ridge regression performed better than modified ridge regression are shown in Table 5.

Table 5. Number of sets where ordinary ridge regression performed better than modified ridge regression in 12 sets

Group n	D41 25	D41 75	D42 25	D42 75	D81 25	D81 75	D82 25	D82 75
$\rho_{y\hat{y}/\underline{b}}$	9	10	9	11	7	9	8	9
$\varepsilon(y-\hat{y})^2/\underline{b}$	9	11	9	11	7	8	7	7

Modified ridge regression requires extra work to estimate measurement error variance. Therefore, in order to justify usage of modified ridge regression over ordinary ridge regression, modified ridge regression must perform much better than ridge regression, i.e., result in smaller numbers in Table 5. The test to compare modified ridge regression with ordinary ridge regression is one-tail test of the null hypothesis that modified ridge regression is equal to ordinary ridge regression against the alternative hypothesis that modified ridge regression is better. However, all entries

of Table 5 are larger than 6, which is half of the 12 sets. Therefore, the present research suggested that the extra effort to estimate measurement error variances is not justified. The failure of modified ridge regression might be due to the process used to estimate measurement error variances. In order to estimate measurement error variances accurately, larger sample sizes than those used in the present research may be required. However, as mentioned in Chapter II, when the sample size increases, performance of the least square method will get better. Therefore, the method of modified ridge regression may have a self-defeating nature. Another possible reason for modified ridge regression not to have performed well may be due to values of k. Since values of k for (4.17) were given by dividing k values of (3.7) by one minus average reliabilities of predictor variables, intervals between adjacent values of k were much greater for modified ridge regression than for ordinary ridge regression. Therefore, there may be better values of $\rho_{y\hat{y}/\underline{b}}$ and $\varepsilon(y-\hat{y})^2/\underline{b}$ for modified ridge regression than those values obtained in the present research where the optimum k value lies somewhere between an adjacent pair of k values used in this research.

The number of sets where the best values of two criteria for ordinary ridge regression performed better than least square are shown in Table 6.

Table 6. Number of sets where ordinary ridge regression with the best k-value performed better than least square in 12 sets

Group n	D41 25	D41 75	D42 25	D42 75	D81 25	D81 75	D82 25	D82 75
$\rho_{y\hat{y}/\underline{b}}$	10*	8	8	3	12***	11**	12***	11**
$\varepsilon(y-\underline{y})^2/\underline{b}$	11**	7	10*	5	12***	10*	12***	10*

* $p<0.019$

** $p<0.0032$

*** $p<0.00024$

The Binomial distribution was used to test the null hypothesis that ordinary ridge regression was equal to least square against the alternative hypothesis that ordinary ridge regression was better than least square. Table 6 shows that when n=25 ordinary ridge regression performed significantly better than least square except for one cell, and when the number of predictor variables was 8 ordinary ridge regression performed better than least square. This result was consistent with the results of Chapter II which showed that goodness of prediction by least square was inversely related to the number of predictors and positively related to sample size.

In order to find the extent to which ordinary ridge regression was better than least square, best values of

criteria obtained by ordinary ridge regression were plotted against values of criteria obtained by least square in Figure 1 through Figure 8. All figures, except for two $\rho_{y\hat{y}/\underline{b}}$'s of sample size 75 with four predictor variables, showed that when least square performed well, ordinary ridge regression performed as well as least square, i.e., distances of points from the straight lines were small, whereas when least square did not perform well, ordinary ridge regression performed much better, i.e., distances of points from the straight lines were large. Generally, therefore, ordinary ridge regression had small risk when it performed worse than least square and it had a large advantage when it performed better than least square.

Means, standard deviations (SD's) and ranges of k values for ordinary ridge regression used in Table 5, Table 6 and Figure 1 through Figure 8 are shown in Table 7. Table 7 shows that the best k value for $\rho_{y\hat{y}/\underline{b}}$ and the best k value for $\varepsilon(y-\hat{y})^2/\underline{b}$ of a given set of data were not necessarily the same. This means that for different criteria, the optimal k may be different for a given sample variance matrix. This result was not surprising because the intercept and scale have no influence on $\rho_{y\hat{y}/\underline{b}}$ but do influence $\varepsilon(y-y)^2/\underline{b}$.

Table 7 also shows that optimum k values are large when the sample size was small, or when the number of predictor

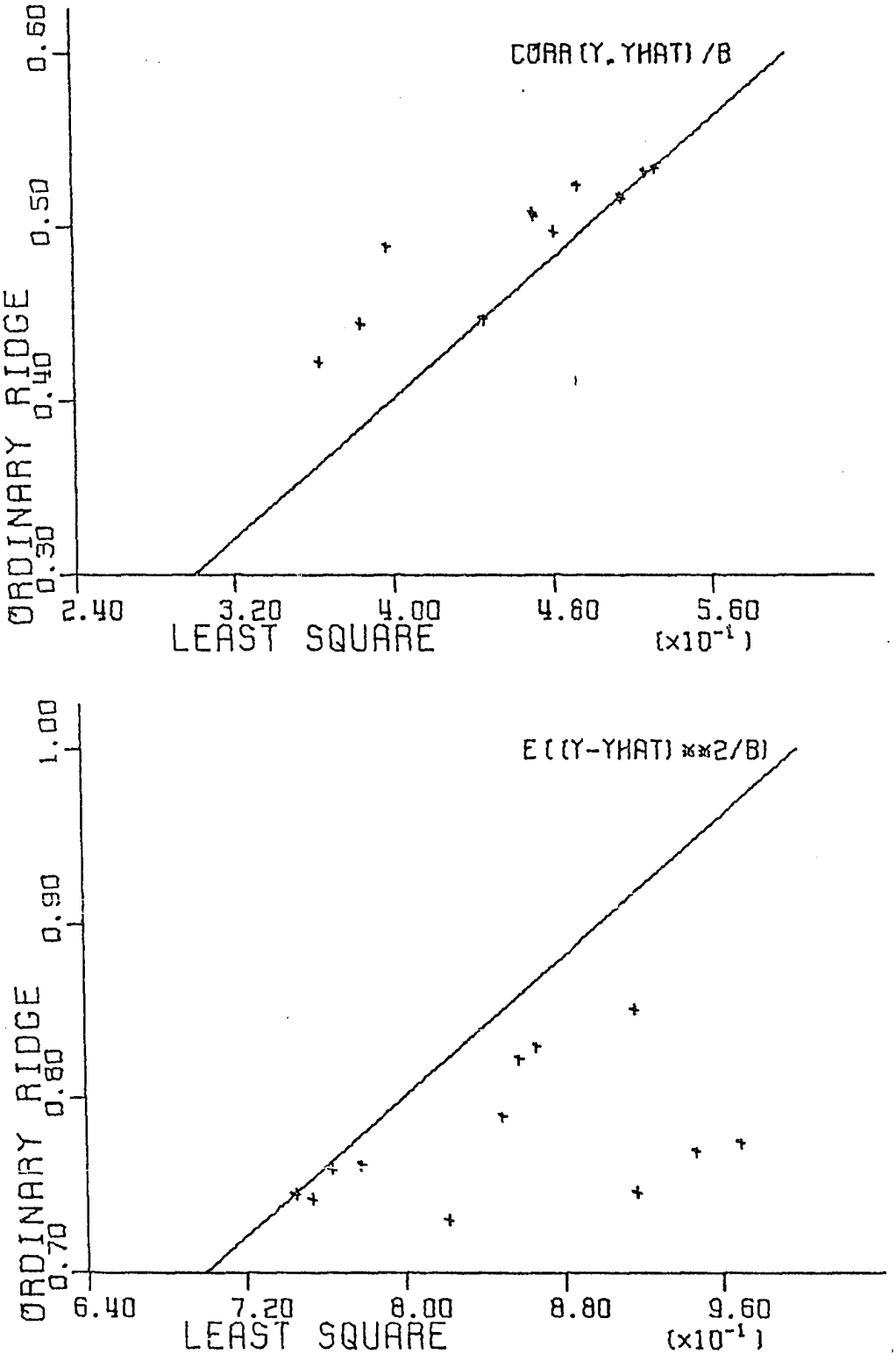

FIGURE 1. ORDINARY RIDGE WITH THE BEST K VALUE VS LS ON TWO CRITERIA. GROUP=041, N=25.

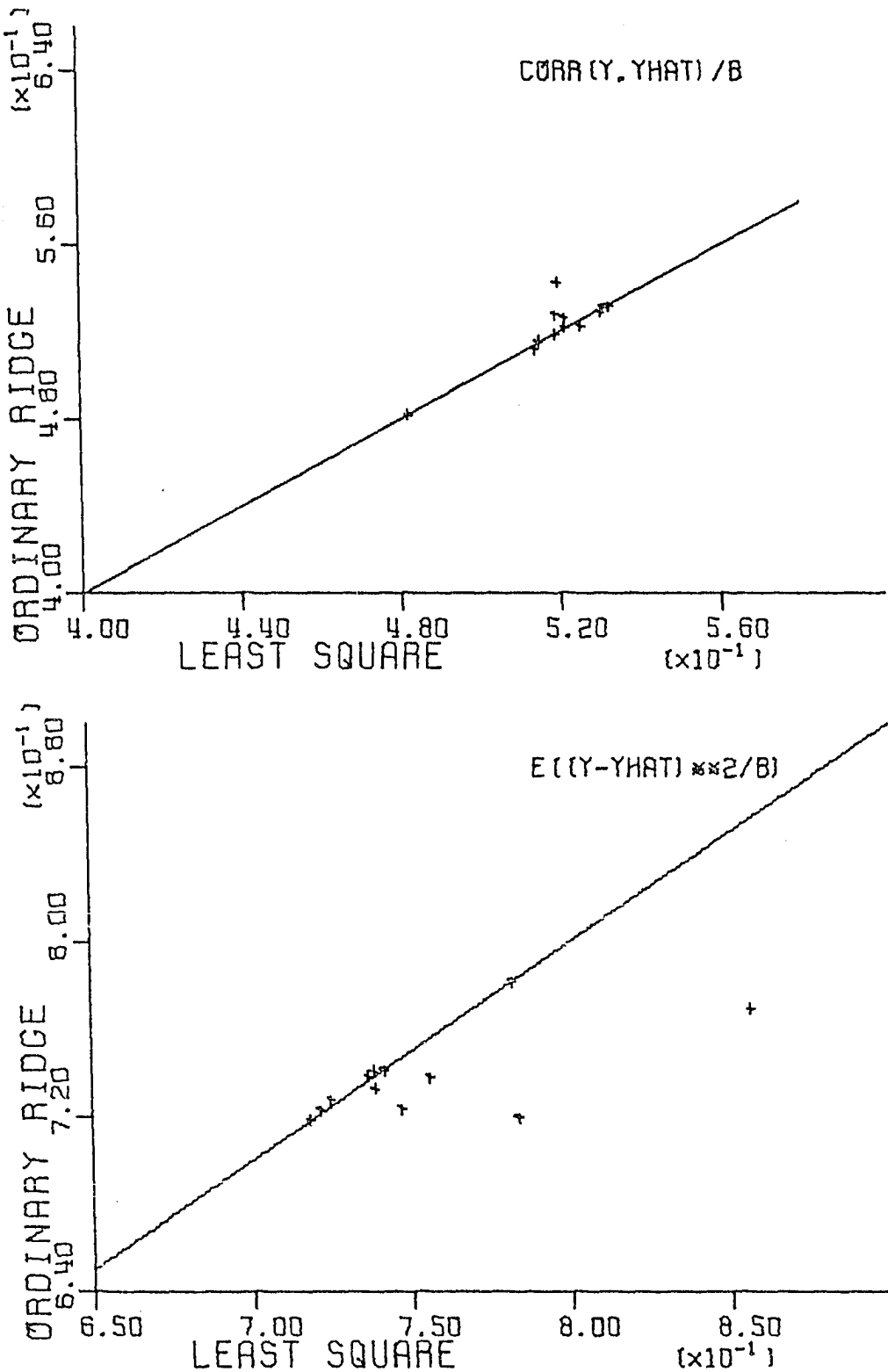

FIGURE 2. ORDINARY RIDGE WITH THE BEST K VALUE VS LS ON TWO CRITERIA. GROUP=041. N=75.

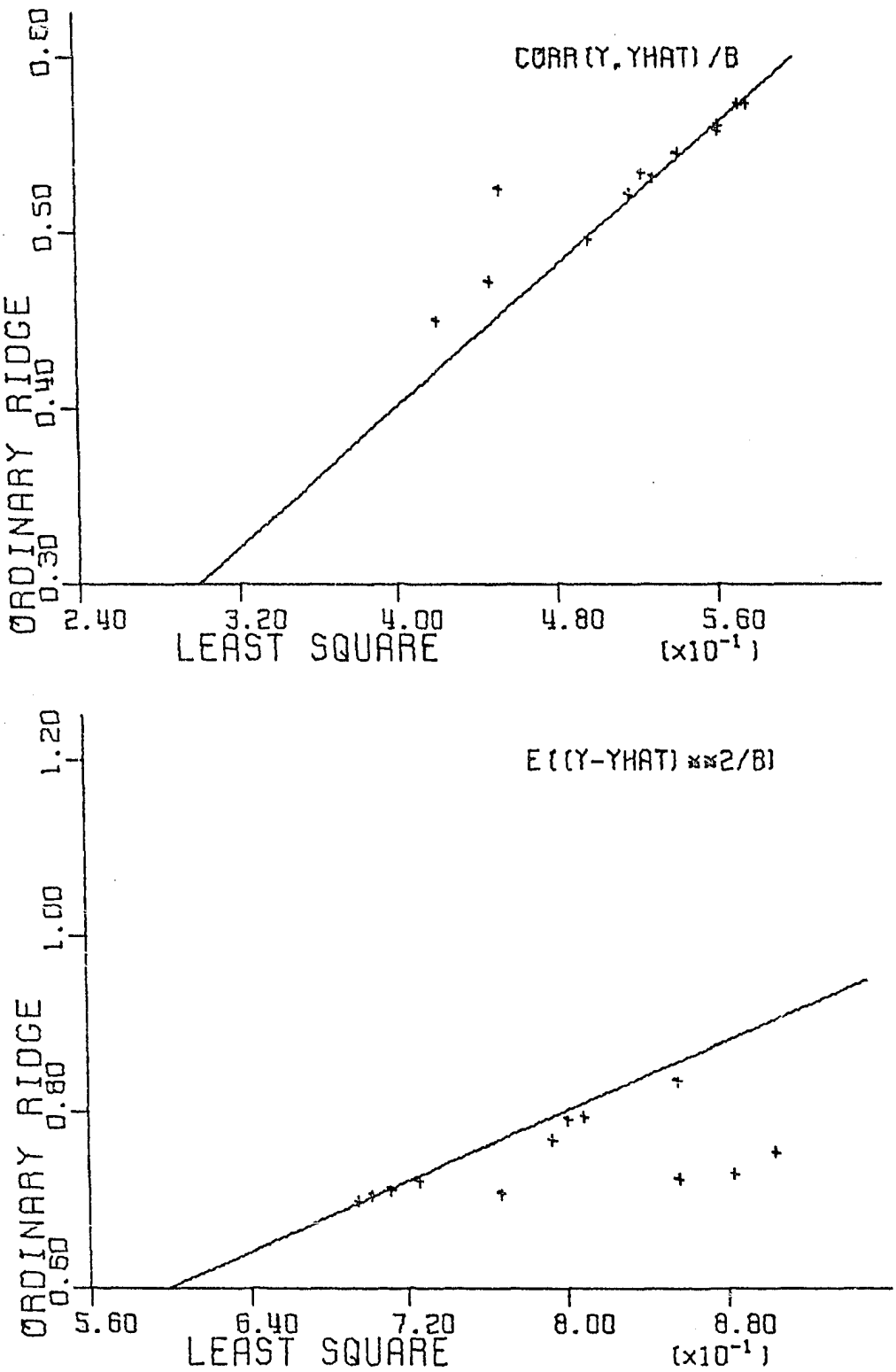

FIGURE 3. ORDINARY RIDGE WITH THE BEST K VALUE VS LS ON TWO CRITERIA. GROUP=D42, N=25.

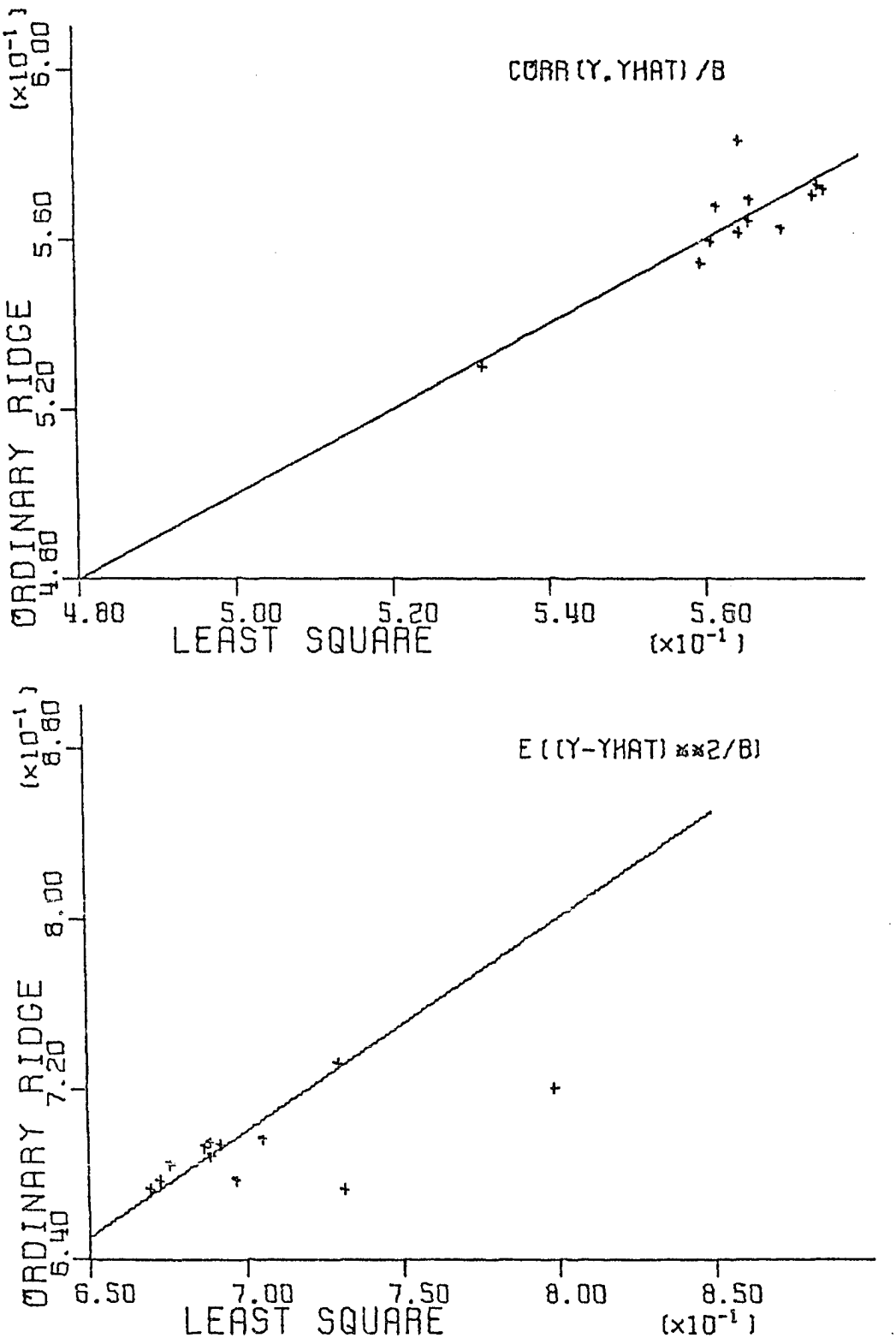

FIGURE 4. ORDINARY RIDGE WITH THE BEST K VALUE VS LS ON TWO CRITERIA. GROUP=D42. N=75.

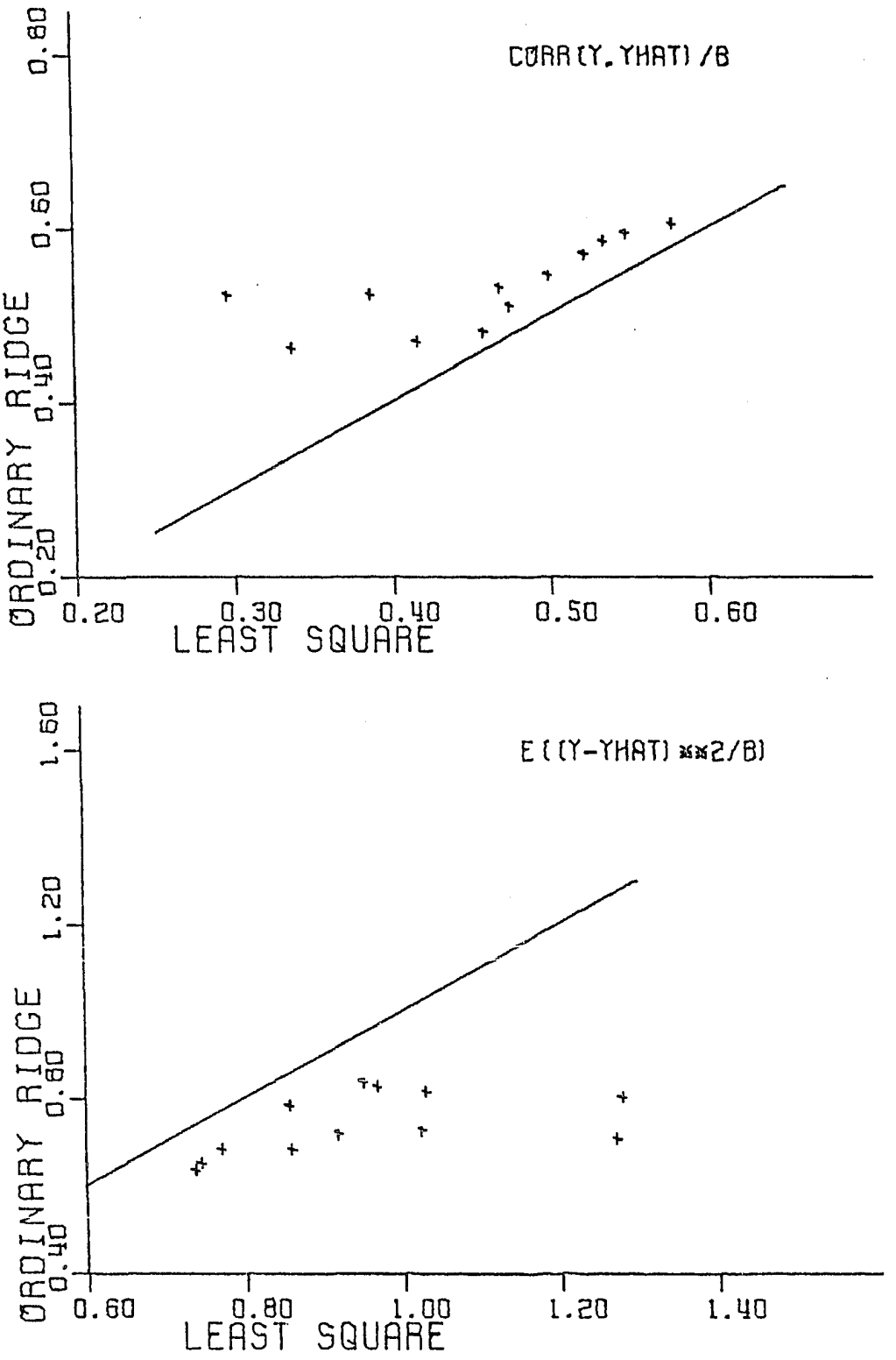

FIGURE 5. ORDINARY RIDGE WITH THE BEST K VALUE VS LS ON TWO CRITERIA. GROUP=D81, N=25.

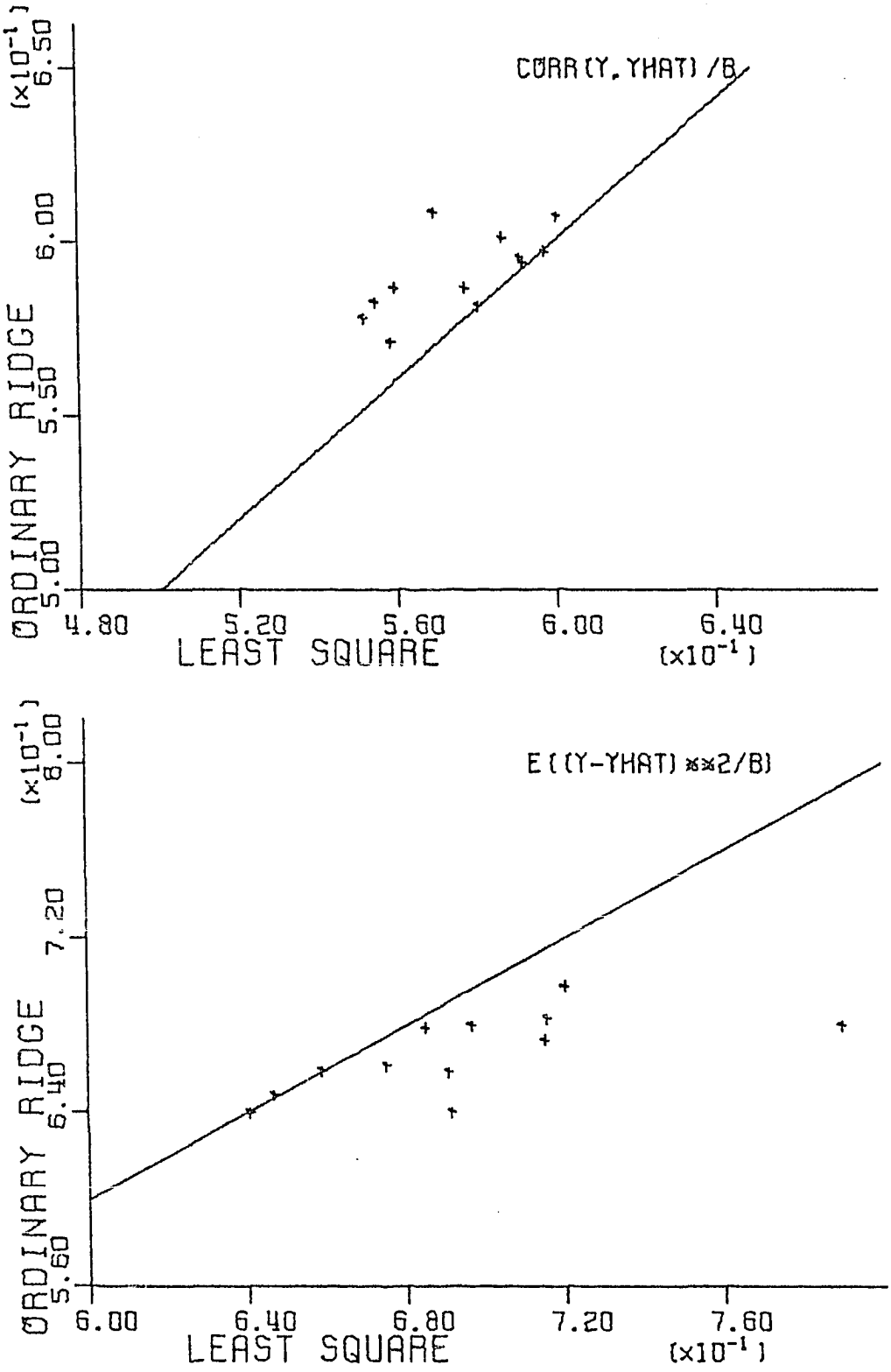

FIGURE 6. ORDINARY RIDGE WITH THE BEST K VALUE VS LS ON TWO CRITERIA. GROUP=081. N=75.

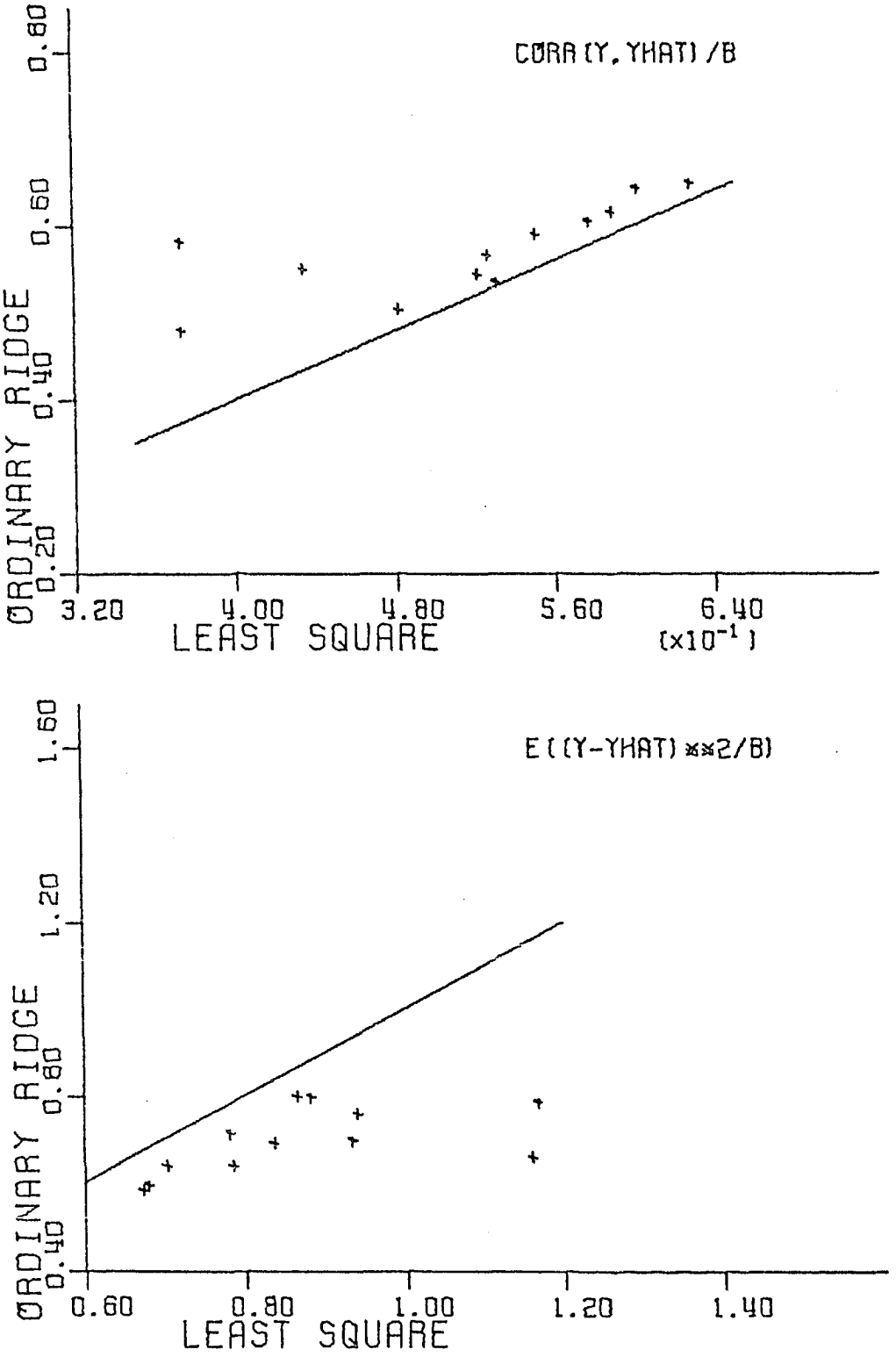

FIGURE 7. ORDINARY RIDGE WITH THE BEST K VALUE VS LS ON TWO CRITERIA. GROUP=D82, N=25.

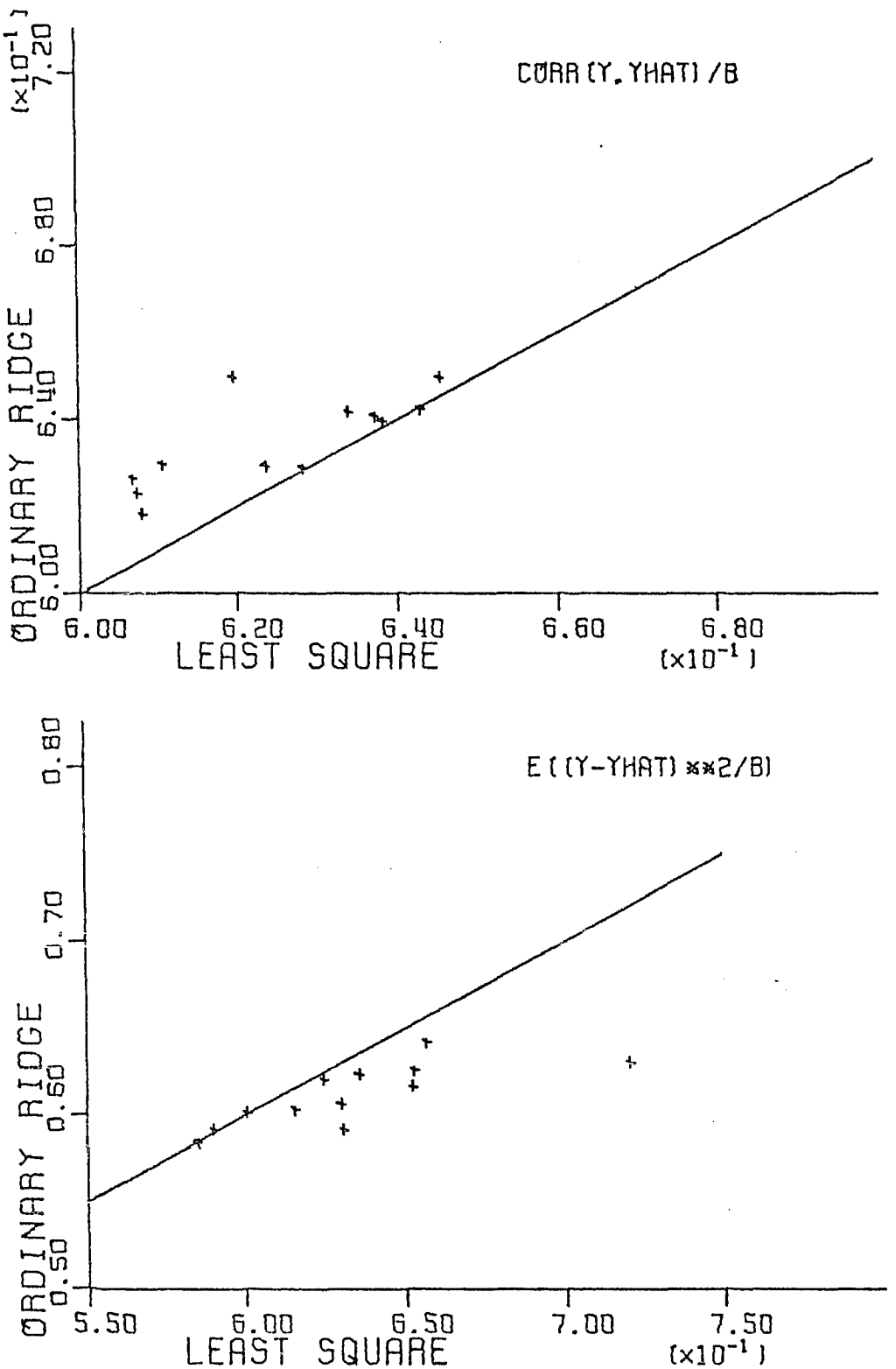

FIGURE 8. ORDINARY RIDGE WITH THE BEST K VALUE VS LS ON TWO CRITERIA. GROUP=D82. N=75.

Table 7. Means, standard deviations and ranges of the best k values of ordinary ridge regression for 12 sets of data

	Group	D41		D42		D81		D82	
	n	25	75	25	75	25	75	25	75
$\rho_{y\hat{y}/\underline{b}}$	mean	.4025	.0225	.1817	.0142	.6983	.1725	.4033	.0983
	SD	.4170	.0242	.3427	.0090	.3197	.1360	.3176	.0995
	range	.99	.07	.99	.03	.92	.39	.96	.29
$\varepsilon(y-\hat{y})^2/\underline{b}$	mean	.3542	.0508	.1900	.0183	.5083	.1350	.3267	.0700
	SD	.2477	.0595	.2061	.0159	.3370	.1115	.2839	.0520
	range	.59	.19	.59	.05	.90	.39	.94	.19

variables was large. This tendency was expected from Chapter II, since least square tends to perform poorly when the sample size is small relative to the number of predictor variables. Measurement errors also seemed to have influence on the values of k. Groups D42 and D82 had variance matrices of predictor variables with smaller diagonal elements than Groups D41 and D81 because of less measurement error. This resulted in less variance of lack of fit of least square. Ranges and standard deviations in Table 7 indicate that values of k did not cluster in a narrow band for data sets in the same group except for those with four predictor variables and n=75.

In order to test the comment by Marquardt and Snee (1975) that the selection of k value is not important, and to test effectiveness of visual inspections to decide a value of k,

ridge traces were plotted together with two criteria for each set of data. Ninety-six plots, however, did not show any pattern of where the optimal k values for $\rho_{y\hat{y}/\underline{b}}$ or $\varepsilon(y-\hat{y})^2/\underline{b}$ occurred. Plots failed to show a flat minimum of $\varepsilon(y-\hat{y})^2/\underline{b}$ contradicting the results of Marquardt and Snee (1975). This may be due to the fact that Marquardt and Snee (1975) used the prediction standard deviation based upon one case of real data. Therefore, their result might have two sources of errors, namely sampling of an initial group and sampling of new data, whereas the present study had only one source of error which was sampling of an initial group. Since the purpose of prediction using a linear function is to obtain the best set of weights from an initial sample, it seems to be desirable to have sampling fluctuations only in an initial sample. The difference between Marquardt and Snee (1975) and the present research might also be due to the fact that $\varepsilon(y-\hat{y})^2/\underline{b}$ is more sensitive to change than is the standard deviation. Generally speaking, k values obtained by visual inspection were larger than the optimal values, which supports the finding by Obenchain (1975).

In an applied situation, researchers never know the parameters for the variables involved. Therefore, they can not calculate $\rho_{y\hat{y}/\underline{b}}$ or $\varepsilon(y-\hat{y})^2/\underline{b}$ in order to choose optimal value of k. Even though the present study indicated the

possibility of existence of optimal k values which maximize $\rho_{y\hat{y}/\underline{b}}$ or minimize $\varepsilon(y-\hat{y})^2/\underline{b}$, such k values can not be obtained from a given sample. However, if a researcher wants to improve prediction using ordinary ridge regression, it is not necessary to use optimal k values. It is sufficient to use some k values which tend to result in better prediction than least square. The present study showed that the optimal k value is a function of a sample size, number of predictor variables, and population parameters. Since a sample size and number of predictor variables are known to researchers, some function which is inversely related to a sample size and positively related to number of predictor variables should give a value of k which tends to result in better prediction than least square. In Study 2, two such functions were investigated.

CHAPTER VI. STUDY 2

Method

Population Parameters

Population parameters used in study 2 were the same as those used in study 1.

Factors Manipulated

In Study 2 five main factors and one minor factor were investigated. The first three factors were the same as for Study 1, namely the number of predictor variables, the degree of measurement errors in predictor variables and the sample size. (However for the factor of sample size, the sample size 50 was added.) The fourth factor was restriction in sampling. Observations with values of criterion variable under a cutoff point were deleted from the initial sample. Cutoff points were set to $-\infty$, -1.5, -1.0, -0.5, and 0.0. The cutoff point of $-\infty$ was the case where no deleting of observations occurred. The fifth factor was the procedures used to obtain prediction weights. The least square method, equal weighting based upon the model (3.14) with least square, and ordinary ridge regression were investigated. One minor factor was values of k used for ordinary ridge regression. Since Study 1 suggested that k values given by some functions which were inversely related to a sample size and positively

related to the number of predictor variables, two such functions were used in Study 2. One was given by

$$k = p/n \tag{6.1}$$

and the other was given by

$$k = p/(n-p-1). \tag{6.2}$$

Procedure

Twenty sets of independent vectors above the cutoff point were generated from multivariate normal distributions with mean zero and the variance matrix augmented by the criterion variable with an appropriate sample size. The diagonal elements of the variance matrix for the predictor variables were sums of ones and error variances.

For each set of data, obtained scores were standardized. Then, equation (3.7) was used to obtain weights for least square and ordinary ridge regression by setting k equal to zero and those values given by (6.1) and (6.2). The model (3.14) was used for equal weighting and least square was used for estimation of parameters. Weights were then converted back to the original scale using sample means and standard deviations.

Two criteria used in Study 1 were computed for each set of data using (5.2) and (5.3).

There were four groups of data, namely D41, D42, D81 and D82 which were generated by four different variance

matrices as was the case of Study 1. For each group, three hundreds (20 sets x 3 sample sizes x 5 cutoffs) sets of data were generated.

Results and Discussion

Numbers of sets which performed better than least square were counted for ridge regression with k=p/n (R_1) and k=p/(n-p-1) (R_2) and for equal weighting, and they are shown in Table 8 through Table 11. The Binomial distribution was used to test the null hypothesis that another method was equal to least square against the alternative hypothesis that the other method was better than least square.

Tables 8 through 11 show that equal weighting was not efficient for replacing the least square method for the four variance matrices used in the present study. This might be due to the fact that $(\Sigma_{22}+G)^{-1}\Sigma_{21}$ had negative entries. The result did not support the finding by Schmidt (1971) where unit weighting was superior in correlation when the sample size was equal to 25 and the numbers of predictor variables were 2, 4, 6, 8 and 10 with and without suppressor variables. Laughlin (1978) and Pruzek and Frederick (1978) stated that equal weighting could perform relatively well in certain situations, but other weighting methods could be developed which would lead to more optimally valid and stable prediction. The present research supports their comment.

Table 8. Number of sets out of 20 sets which performed better than least square for Group D41.

Method	n	$\varepsilon(y-\hat{y})^2/\underline{b}$ cutoff $-\infty$	-1.5	-1.0	-0.5	0.0	$\rho_{y\hat{y}}/\underline{b}$ cutoff $-\infty$	-1.5	-1.0	-0.5	0.0
	25	18	16	9	9	7	15	14	10	10	14
R_1[a]	50	11	6	6	3	4	9	12	11	13	13
	75	10	4	5	0	1	10	10	7	12	11
	25	17	15	9	8	7	15	14	10	10	14
R_2[b]	50	11	6	6	3	4	9	11	11	13	13
	75	10	4	5	0	1	9	9	7	13	11
	25	6	8	5	5	5	4	2	5	6	12
equal weight	50	1	0	1	0	1	2	0	3	3	8
	75	1	1	0	0	1	0	0	0	1	2

Entry
18 $p < 0.0002$
17 $p < 0.0013$
16 $p < 0.0059$
15 $p < 0.0207$
14 $p < 0.0575$

[a] R_1 = Ridge regression with k=p/n.

[b] R_2 = Ridge regression with k=p/(n-p-1).

Table 9. Number of sets out of 20 sets which performed better than least square for Group D42.

		$\varepsilon(y-\hat{y})^2/\underline{b}$					$\rho_{y\hat{y}}/\underline{b}$				
		cuttoff					cutoff				
Method	n	$-\infty$	-1.5	-1.0	-0.5	0.0	$-\infty$	-1.5	-1.0	-0.5	0.0
R_1	25	12	12	4	2	5	9	8	4	5	7
	50	7	3	3	0	0	4	5	6	6	10
	75	6	1	1	0	1	2	5	3	6	6
R_2	25	12	11	4	2	5	8	8	4	5	7
	50	7	3	3	0	0	4	4	6	6	9
	75	6	1	1	0	1	2	5	3	5	5
Equal Weight	25	4	4	3	2	4	2	0	5	5	5
	50	0	1	1	0	0	0	0	2	0	2
	75	0	0	0	0	1	0	0	0	0	1

Table 10. Number of sets out of 20 sets which performed better than least square for Group D81

		$\varepsilon(y-\hat{y})^2/\underline{b}$					$\rho_{y\hat{y}}/\underline{b}$				
		cutoff					cutoff				
Method	n	$-\infty$	-1.5	-1.0	-0.5	0.0	$-\infty$	-1.5	-1.0	-0.5	0.0
	25	18	18	16	17	6	16	19	17	19	15
R_1	50	19	12	10	6	1	17	13	18	12	18
	75	17	10	9	2	1	16	10	14	15	16
	25	17	17	15	12	5	14	19	16	17	15
R_2	50	19	11	9	6	1	16	12	18	12	18
	75	15	10	7	2	0	16	9	14	15	16
	25	14	10	8	6	4	6	7	9	14	10
Equal Weight	50	3	1	0	3	1	1	2	2	4	3
	75	1	1	0	0	0	1	1	0	0	5

Entry	
19	$p < 0.0000$
18	$p < 0.0002$
17	$p < 0.0013$
16	$p < 0.0059$
15	$p < 0.0207$
14	$p < 0.0575$

Table 11. Number of sets out of 20 sets which performed better than least square for Group D82

		$\varepsilon(y-\hat{y})^2/\underline{b}$					$\rho_{y\hat{y}}/\underline{b}$				
		cutoff					cutoff				
Method	n	$-\infty$	-1.5	-1.0	-0.5	0.0	$-\infty$	-1.5	-1.0	-0.5	0.0
	25	14	17	14	9	3	12	16	13	17	14
R_1	50	15	8	4	3	2	13	10	11	13	17
	75	12	4	2	1	0	11	6	12	11	13
	25	14	14	13	8	3	11	15	13	18	13
R_2	50	14	8	1	3	2	11	10	11	13	14
	75	12	4	2	0	0	11	6	12	11	12
	25	7	11	7	5	1	3	6	8	5	9
Equal Weight	50	1	0	0	0	1	1	0	0	1	2
	75	1	0	0	0	0	0	0	0	0	1

Entry	
18	$p < 0.0002$
17	$p < 0.0013$
16	$p < 0.0059$
15	$p < 0.0207$
14	$p < 0.0575$

Tables 8, 10 and 11 show that least square was likely to perform worse than ridge regression when the sample size was small and/or the number of predictors was large. However, population variance matrices were also important factors. Groups D42 and D82 had smaller variances of predictor variables because of less measurement errors than groups D41 and D81. This increased the ratio given by $\lambda_{max}/\lambda_{min}$, which is supposed to favor ridge regression. However, the decrease in variances of predictor variables reduced the residual variance which seemed to favor least square, more than offsetting the increase in $\lambda_{max}/\lambda_{min}$.

Means ($\bar{x}$'s) and variances (s^2's) of $\rho_{y\hat{y}/\underline{b}}$ and $\varepsilon(y-\hat{y})^2/\underline{b}$ were calculated for least square (LS) and ridge regression with different k values (R_1 and R_2) and shown in Table 12 through Table 15.

Table 8 through Table 15 indicate that there was not much difference between R_1 and R_2 in the present research. Table 8 through Table 11 show that R_1 more often performed better than least square compared with R_2. Yet, Table 12 through Table 15 show that $\bar{x}$'s and s^2's of R_1 and R_2 were very close in the cells where Table 8 through Table 11 show differences. The present research was not conclusive about k values used here. Tentatively k=p/n is recommended because of greater frequency to perform better than least square under certain conditions.

Table 12. $\bar{x}$ and s^2 of 20 sets on two criteria for Group D41.

			$\varepsilon(y-\hat{y})^2/\underline{b}$					$\rho_{y\hat{y}}/\underline{b}$				
			cutoff					cutoff				
	n	Method	$-\infty$	-1.5	-1.0	-0.5	0.0	$-\infty$	-1.5	-1.0	-0.5	0.0
		LS	887	871	914	1114	1390	452	475	422	352	332
	25	R_1	821	818	894	1116	1405	480	496	418	355	361
		R_2	821	819	896	1118	1409	480	496	417	355	362
		LS	777	786	857	973	1339	510	501	472	480	442
$\bar{x}$	50	R_1	770	790	861	994	1365	510	505	480	485	459
(10^{-3})		R_2	770	790	862	995	1368	510	505	480	485	460
		LS	748	770	820	970	1299	521	508	506	507	480
	75	R_1	744	772	828	990	1320	521	511	505	511	485
		R_2	744	772	828	991	1322	521	511	505	511	485

		LS	98	197	110	641	807	40	20	140	433	563
	25	R_1	36	88	105	507	723	15	13	210	586	684
		R_2	35	84	104	490	710	13	12	218	604	700
		LS	26	18	74	94	258	14	10	53	55	83
s^2	50	R_1	17	15	60	99	260	10	9	43	48	57
(10^{-4})		R_2	17	15	59	99	260	10	9	42	48	55
		LS	11	17	17	63	134	3	10	10	13	75
	75	R_1	7	10	16	58	103	3	4	7	13	56
		R_2	7	10	16	57	102	3	4	7	13	55

$R_{y\hat{y}} = 545.5 \times 10^{-3}$

$1 - R^2_{y\hat{y}} = 702.4 \times 10^{-3}$

Table 13. $\bar{x}$ and s^2 of 20 sets on two criteria for Group D42

			$\varepsilon(y-\hat{y})^2/\underline{b}$					$\rho_{y\hat{y}}/\underline{b}$				
			cutoff					cutoff				
	n	Method	$-\infty$	-1.5	-1.0	-0.5	0.0	$-\infty$	-1.5	-1.0	-0.5	0.0
		LS	827	825	839	1041	1312	504	521	491	439	390
	25	R_1	794	802	874	1088	1389	508	521	464	416	378
		R_2	798	807	880	1094	1396	506	518	462	414	376
		LS	725	748	803	905	1257	556	542	523	539	501
$\bar{x}$	50	R_1	735	772	832	960	1325	542	533	515	532	501
(10^{-3})		R_2	737	774	834	964	1329	540	532	514	530	500
		LS	701	712	764	911	1232	567	559	556	547	526
	75	R_1	709	733	792	959	1285	556	552	547	539	523
		R_2	710	734	794	960	1287	555	551	546	538	522

s^2 (10^{-4})											
		LS	86	216	98	578	860	32	20	61	185	523
	25	R_1	35	94	85	496	681	13	24	71	317	644
		R_2	35	89	86	485	655	12	23	69	322	666
		LS	23	28	83	70	298	11	14	51	31	38
	50	R_1	17	22	64	83	292	10	11	38	29	38
		R_2	16	22	63	84	291	10	11	37	29	37
		LS	11	9	27	51	139	2	4	10	8	75
	75	R_1	7	7	24	45	94	2	5	7	6	61
		R_2	7	7	24	45	93	2	5	7	6	60

$R_{y\hat{y}} = 587.1 \times 10^{-3}$

$1 - R^2_{y\hat{y}} = 655.3 \times 10^{-3}$

Table 14. $\bar{x}$ and s^2 of 20 sets on two criteria for Group D81

			$\varepsilon(y-\hat{y})^2/\underline{b}$					$\rho_{y\hat{y}}/\underline{b}$				
			cutoff					cutoff				
	n	Method	$-\infty$	-1.5	-1.0	-0.5	0.0	$-\infty$	-1.5	-1.0	-0.5	0.0
		LS	938	989	893	1076	1313	478	414	441	412	351
	25	R_1	764	815	824	1027	1350	519	504	481	487	375
		R_2	761	818	831	1039	1367	521	510	483	495	375
		LS	779	758	766	938	1196	520	523	533	503	508
$\bar{\bar{x}}$	50	R_1	728	744	763	954	1238	543	535	550	525	531
(10^{-3})		R_2	727	746	767	959	1247	543	535	551	526	532
		LS	718	697	746	893	1182	559	560	553	540	501
	75	R_1	689	697	753	912	1223	570	566	560	551	516
		R_2	689	699	755	914	1228	571	566	560	552	517

s^2											
		LS	452	537	246	326	527	75	207	177	145	326
	25	R_1	30	91	103	200	398	24	73	131	54	469
		R_2	23	71	92	205	380	20	55	116	45	531
		LS	64	34	38	132	193	24	21	30	73	34
	50	R_1	29	30	28	89	161	15	19	19	36	23
		R_2	27	31	28	87	156	14	19	18	33	21
		LS	31	23	10	49	137	17	18	4	8	52
	75	R_1	10	15	10	42	118	7	8	5	5	39
		R_2	9	15	10	42	116	7	8	5	5	38

$R_{y\hat{y}} = 614.2 \times 10^{-3}$

$1 - R^2_{y\hat{y}} = 588.9 \times 10^{-3}$

Table 15. $\bar{x}$ and s^2 of 20 sets on two criteria for Group D82

			$\varepsilon(y-\hat{y})^2/\underline{b}$					$\rho_{y\hat{y}}/\underline{b}$				
			cutoff					cutoff				
	n	Method	$-\infty$	-1.5	-1.0	-0.5	0.0	$-\infty$	-1.5	-1.0	-0.5	0.0
		LS	856	918	840	995	1189	535	474	489	478	415
	25	R_1	722	779	789	982	1284	559	538	518	532	432
		R_2	729	787	806	1001	1310	555	539	514	533	425
		LS	711	692	710	842	1112	574	581	581	571	553
$\bar{x}$	50	R_1	677	696	727	884	1185	588	582	592	582	568
(10^{-3})		R_2	680	701	734	892	1197	587	581	591	580	568
		LS	655	637	685	821	1094	608	615	601	592	572
	75	R_1	637	650	704	860	1171	614	613	604	598	580
		R_2	638	653	707	865	1179	613	612	603	598	580

		LS	376	491	161	462	354	61	205	127	115	210
	25	R_1	28	99	86	235	273	20	65	115	42	313
		R_2	23	73	76	228	274	18	51	108	36	368
		LS	54	23	33	86	145	20	14	22	25	24
s^2	50	R_1	24	27	28	66	126	12	13	16	10	22
(10^{-4})		R_2	22	27	28	65	123	11	13	15	10	22
		LS	25	19	11	65	125	14	11	5	14	44
	75	R_1	9	14	13	49	123	6	5	6	7	36
		R_2	9	14	13	48	123	5	5	6	6	35

$R_{y\hat{y}} = 657.1 \times 10^{-3}$

$1 - R^2_{y\hat{y}} = 568.2 \times 10^{-3}$

Tables 8 through 15 show that the two criteria used in the research were different in nature. As mentioned in Chapter I, $\varepsilon(y-\hat{y})^2/\underline{b}$ depends upon scale while $\rho_{y\hat{y}/\underline{b}}$ does not. Tables 12 through 15 show that some $\varepsilon(y-\hat{y})^2/\underline{b}$'s, especially when samples were truncated heavily, were greater than one, even though $\rho_{y\hat{y}/\underline{b}}$'s were positive and relatively high. It should be noted that the variance of the criterion variable was set to one. Therefore, $\varepsilon(y-y)^2/\underline{b}$ greater than one means that weights were not good for an absolute prediction purpose even though they might be effective for relative prediction.

Table 8 and Table 9 suggests that ridge regression was likely to perform better than least square when the sample size was 25 and the cutoff was less than -1.5 except the $\rho_{y\hat{y}/\underline{b}}$ for group D42. This was also supported by Table 12 and Table 13, which show better means and smaller variances. In order to find gains and losses, ridge regression with k=p/n (R_1) was plotted against least square using two criteria for groups D41 and D42 with n=25 and cutoff=-∞ and -1.5. They are shown in Figure 9 through Figure 12. Figure 9 and Figure 10 show that when least square performed well, ridge regression was slightly better or worse than least square, whereas when least square did not perform well ridge regression was likely to be much better than least square. This

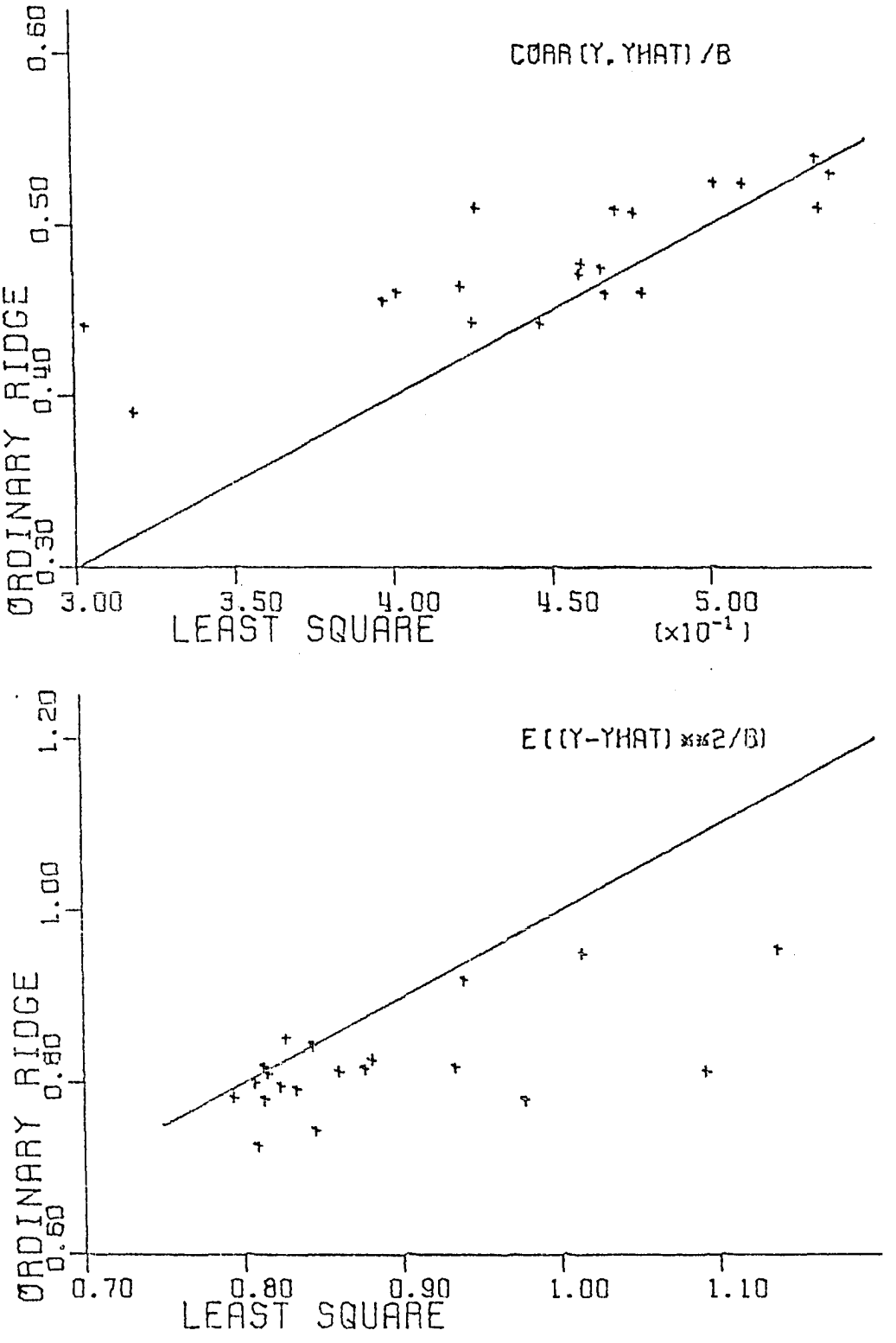

FIGURE 9. ORDINARY RIDGE (R1) VS LS ON TWO CRITERIA.
N=25. CUTOFF=NONE. GROUP=041.

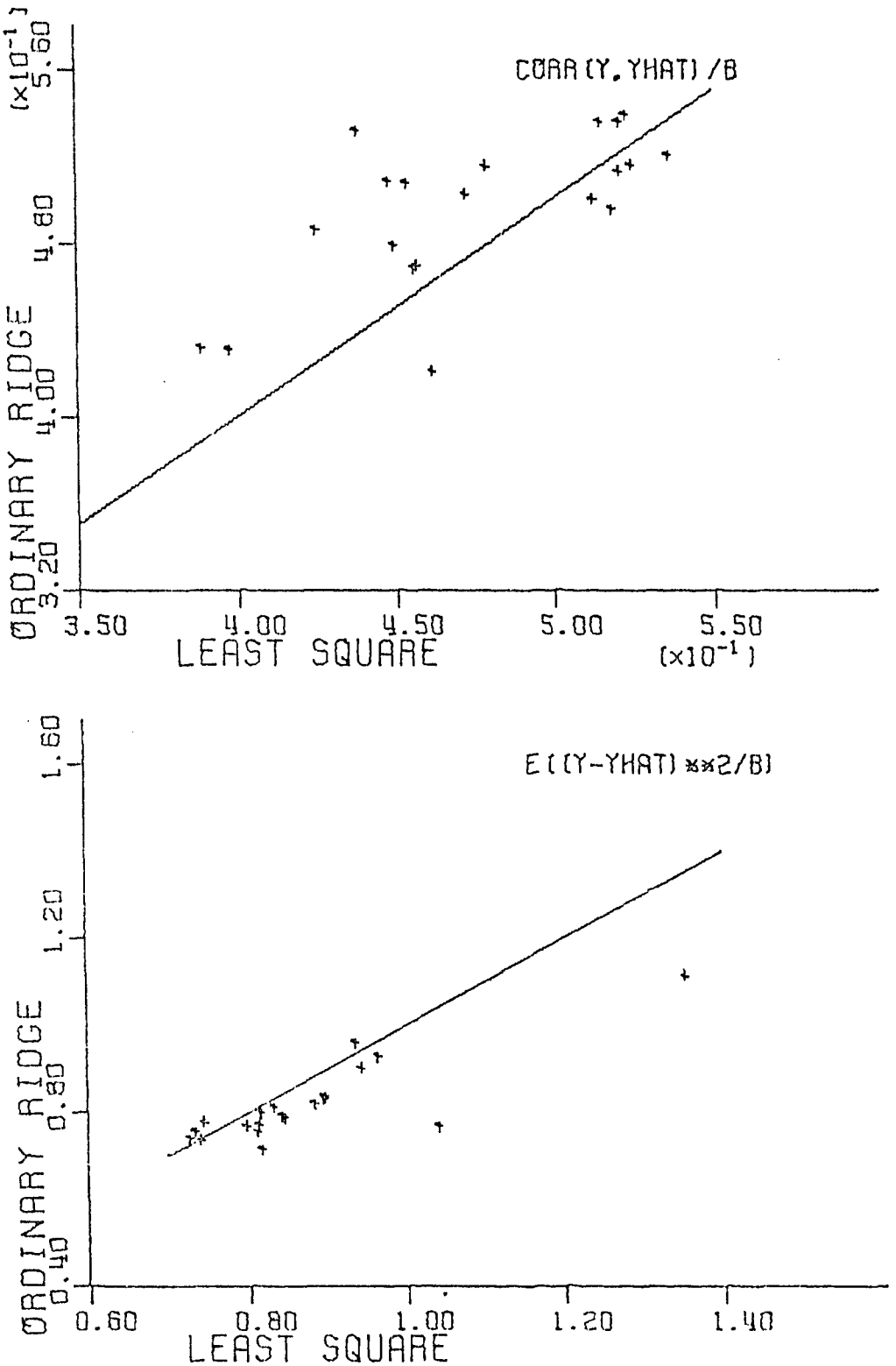

FIGURE 10. ORDINARY RIDGE (R1) VS LS ON TWO CRITERIS. N=25, CUTOFF=-1.5, GROUP=041.

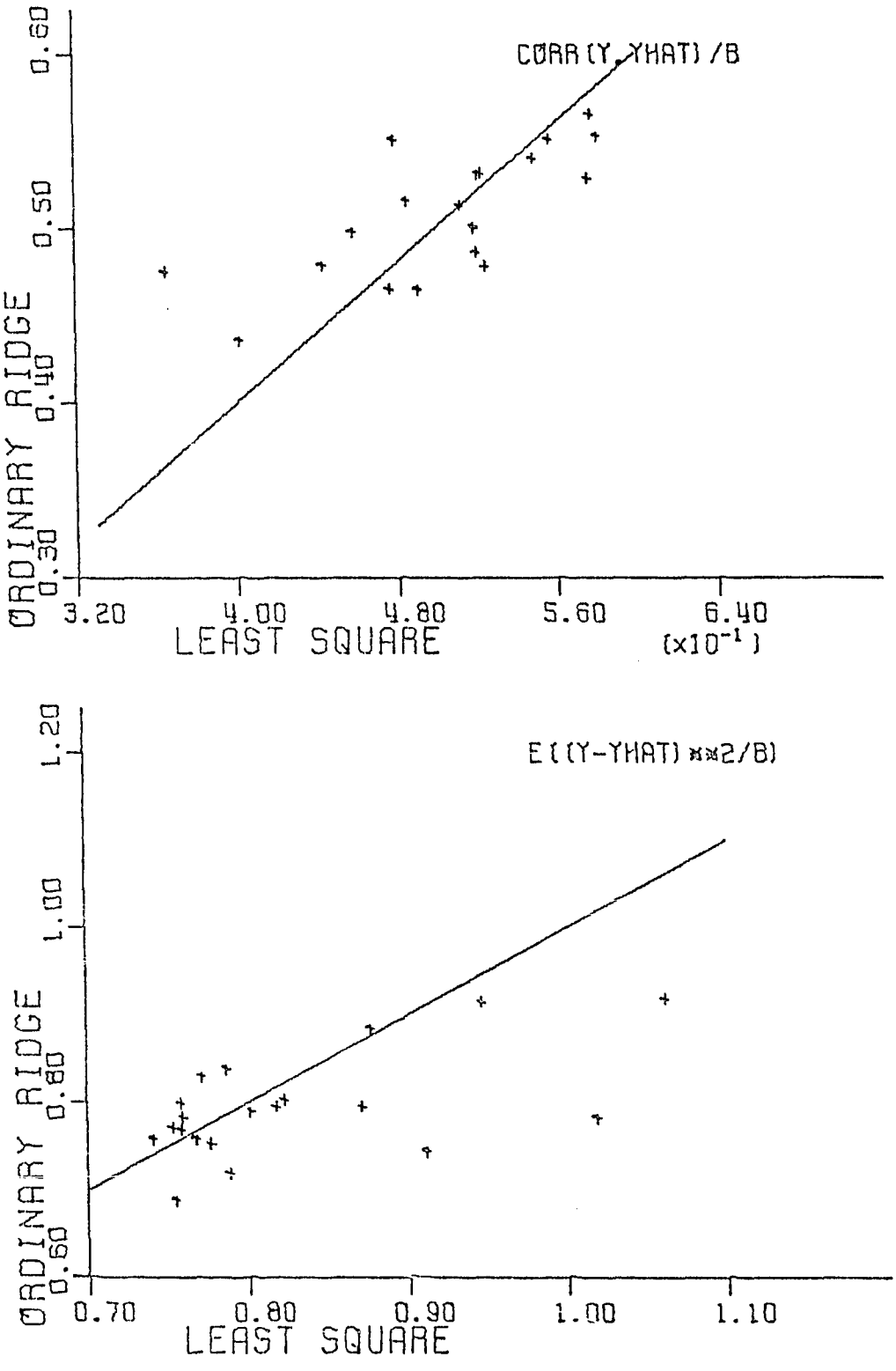

FIGURE 11. ORDINARY RIDGE (R1) VS LS ON TWO CRITERIA. N=25. CUTOFF=NONE. GROUP=D42.

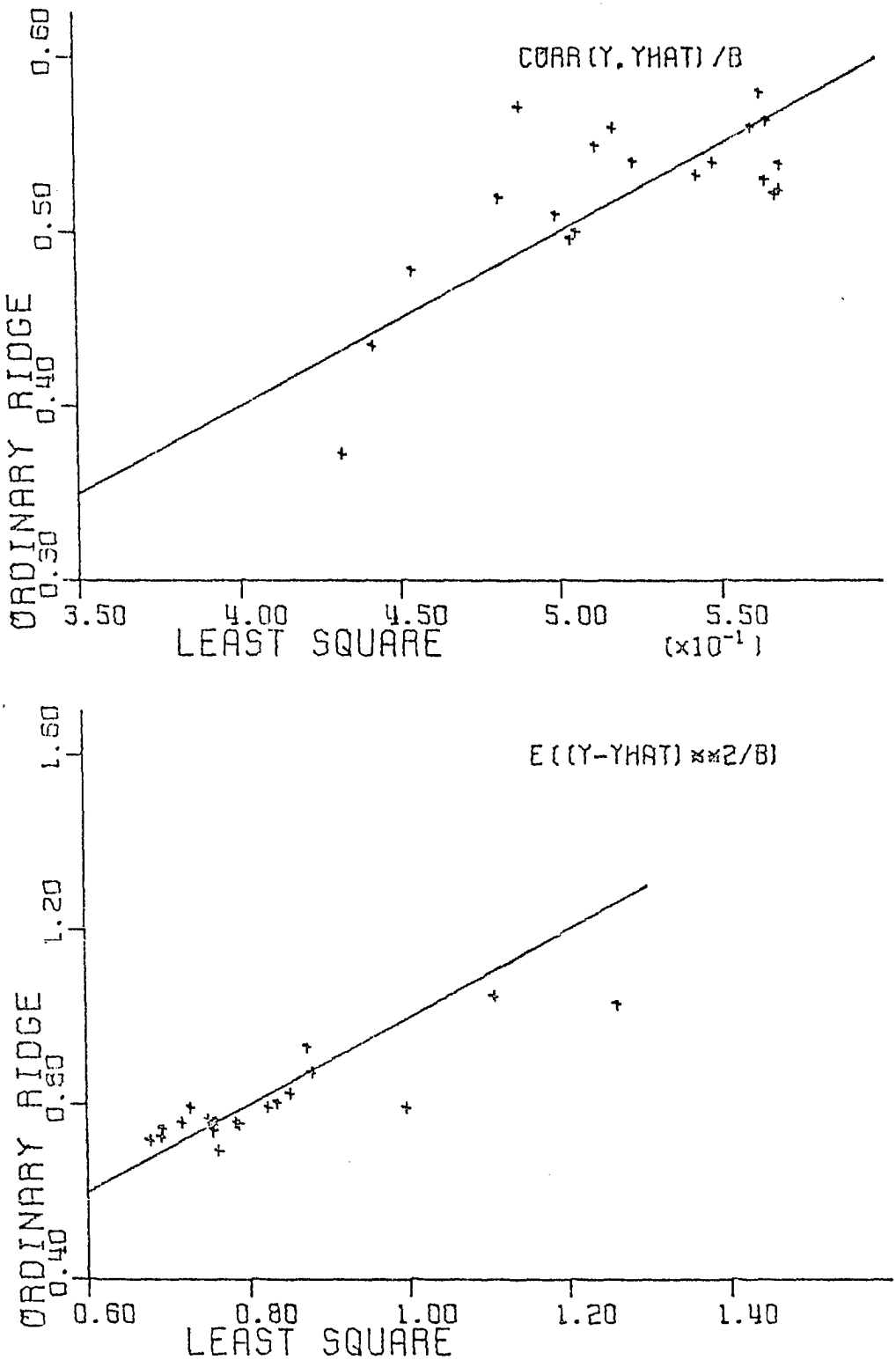

FIGURE 12. ORDINARY RIDGE (R1) VS LS ON TWO CRITERIS.
N=25, CUTOFF=-1.5, GROUP=042.

result was consistent with Study 1. However, Figure 11 and Figure 12 show that for $\rho_{y\hat{y}/\underline{b}}$ this tendency did not hold, and when least square performed well least square was likely to be better than ridge on $\rho_{y\hat{y}/\underline{b}}$. Tables 8, 9, 12 and 13 suggest that other than the cases with n=25 and cutoff less than -1.5 ridge regression with k values used in this research did not have much gain or performed poorer than least square on two criteria.

Tables 10, 11, 13, and 14 show that ridge regression generally performed better than least square on $\rho_{y\hat{y}/\underline{b}}$. However, $\varepsilon(y-y)^2/\underline{b}$ showed that the cutoff point influenced the relative performance of ridge regression. This might be due to the fact that $\varepsilon(y-y)^2/\underline{b}$ was scale dependent. When the samples were truncated, estimates of the mean vector and the variance matrix were biased. Since ridge regression is biased estimation, this additional bias might have severely affected scaling and overcompensated for the instability of least square. When the sample size was 25, general reduction in the values of the weights by ridge regression which stabilized them was more important than this additional bias. However, as the sample size increased, stability of least square increased and biases caused by ridge regression were outweighed. Therefore, ridge regression with 8 predictors was effective for $\varepsilon(y-y)^2/\underline{b}$ when there was no cutoff or less than -1.0 cutoff with n=25.

Values of k used in this study were not the optimal values. Considering the fact that k values fluctuated widely in some conditions (see Table 7), k=p/n and k=p/(n-p-1) had performed relative well to improve prediction in certain conditions.

CHAPTER VII. SUMMARY AND CONCLUSION

Study 1 showed that modified ridge regression did not perform better than ordinary ridge regression. This may be due to the fact that for modified ridge regression it was necessary to estimate the error variances, or due to the fact that the modified ridge regression used more widely spaced k values. Therefore relatively poor performance of modified ridge regression in the present research was not conclusive. Further research is required to investigate those two points before any conclusion about relative performance of modified ridge regression can be drawn.

Study 1 showed that when a sample size was small and/or the number of the predictor variables was large, ridge regression with the optimal k value performed better than least square as shown in Table 6 which was consistent with what was expected from Chapter II. However, optimal k values showed large variability except for the cases with p=4 and n=75 as shown in Table 7. On the average, there exists a k value such that ordinary ridge regression can perform better than least square on two criteria when the sample size is small and/or the number of predictors is large. Figure 1 through Figure 8 show that ridge regression had large gain over least square when it performed better and had small loss when least square performed better. Table 7,

however, shows that an optimal k value for $\rho_{y\hat{y}/\underline{b}}$ and one for $\varepsilon(y-\hat{y})^2/\underline{b}$ are likely to differ.

When ordinary ridge regression is used rather than modified ridge regression, measurement errors do not play any role in the equation except that they are implicitly included in the diagonal. This does not cause any trouble when the same measurement devices are used for the initial sample and the sample where prediction is required, since measurement errors are part of variability inseparable from true variability for those measurement devices.

Study 2 showed that arbitrarily fixed k values, such as k=p/n and k=p/(n-p-1), could perform better than least square under certain conditions. However, the relative performances of ridge regression on two criteria were different. In eight-predictor cases, ridge regression generally performed better for $\rho_{y\hat{y}/\underline{b}}$ than least square as shown in Tables 10 and 11 and ridge regression had higher means and smaller variances of $\rho_{y\hat{y}/\underline{b}}$'s than least square as shown in Tables 14 and 15. However, ridge regression performed better on $\varepsilon(y-y)^2/\underline{b}$ than least square under limited conditions such as no cutoff or n=25 and cutoff less than -1.0 as shown in Tables 10, 11, 14, and 15. Relatively poor performance of ridge regression on truncated data might be due to the fact that ridge regression added biases to existing biases caused by truncation which greatly affected the scale. In the cases

of four predictors, ridge performed better than least square when the sample size was 25 and the cutoff was less than -1.5 for $\varepsilon(y-\hat{y})^2/\underline{b}$ and when the sample size was 25 and there was no cutoff for $\rho_{y\hat{y}/\underline{b}}$.

Study 2 showed that $\varepsilon(y-\hat{y})^2/\underline{b}$ could be larger than the variance of a criterion variable when an initial sample was truncated even though $\rho_{y\hat{y}/\underline{b}}$ was positive and large. This means that the prediction value using weights derived from least square or ridge regression had more squared deviation from an observed value than the population mean μ_y would have. The squared residual $\varepsilon(y-y)^2/\underline{b}$, has been neglected in psychological and educational researches. When cross validity has been investigated, the correlation between observed values and predicted values has been reported. However, present research suggested that some index of reduction in mean squared prediction residual over variance is necessary for absolute prediction validation. Since population parameters are unknown in practice, the following index is proposed as an absolute prediction validity,

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (7.1)$$

where y_i is an observation in a cross validation sample, $\hat{y}_i$ is a predicted value using weights derived from an initial sample, and $\bar{y}$ is a mean in a cross validation sample.

Study 2 showed that equal weighting was not effective and worse than least square most of the time for the variance matrices used here. This might be due to the fact that $(\Sigma_{22} + G)^{-1}\Sigma_{21}$ had negative entries.

Since these results are conditional on the arbitarily selected parameters, no absolute recommendations can be made. However, the results of the present researches suggest the following recommendations to obtain weights:

1. When the number of predictor variable is relatively small and the sample size is relatively small, use ridge regression with k=p/n if there is no truncation in the initial sample, otherwise use least square.
2. When the number of predictor variables is large and the sample size is relatively small or medium, use ridge regression with k=p/n for absolute prediction if there is no truncation in the initial sample and for relative prediction if truncation is less than half of the range of a criterion variable; otherwise use least square.
3. When the sample is severely truncated, seek other methods to obtain weights for absolute prediction.

These recommendations are consistent with what was expected from least square as discussed in Chapter II. However the problem of deciding a good k value is not solved. The value

k=p/n is tentative and further research on values of k is recommended.

REFERENCES

Anderson, T. W. An Introduction to Multivariate Statistical Analysis. New York: John Wiley & Sons, Inc., 1958.

Ayers, J. B. Predicting Quality Point Averages in Master's Degree Programs in Education. Educational and Psychological Measurement, 1971, 31, 491-495.

Dawes, R. M., & Corrigan, B. Linear models in decision making. Psychological Bulletin, 1974, 81, 95-106.

Draper, N. R., & Smith, H. Applied Regression Analysis. New York: John Wiley & Sons, Inc., 1966.

Einhorn, H. J., & Hogarth, R. M. Unit weighting schemes for decision making. Organizational Behavior and Human Performance, 1975, 13, 171-192.

Gorman, J. W., & Toman, R. J. Selection of Variables for Fitting Equations to Data. Technometrics, 1966, 8, 27-51.

Graybill, F. A. Theory and Application of the Linear Model. North Scituate, Massachusetts: Duxbury Press, 1976.

Green, F. G. Parameter sensitivity in multivariate methods. Multivariate Behavioral Research, 1977, 12, 263-288.

Guilkey, D. K., & Murphy, J. L. Directed ridge regression techniques in cases of multicollinearity. Journal of the American Statistical Association, 1975, 70, 769-775.

Gulliksen, H. Theory of Mental Tests. New York: Wiley, 1950.

Hawkins, D. M. Relations Between Ridge Regression and Eigenanalysis of the Augmented Correlation Matrix. Technometrics, 1975, 17, 477-486.

Hemmerle, W. J. An Explicit Solution for Generalized Ridge Regression. Technometrics, 1975, 17, 309-314.

Hocking, R. R. The Analysis and Selection of Variables in Linear Regression. Biometrics, 1976, 32, 1-49.

Hoerl, A. E. Application of ridge analysis to regression problems. Chemical Engineering Progress, 1962, 58, 54-59.

Hoerl, A. E., & Kennard, R. W. Ridge regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 1970, 12, 55-67. (a).

Hoerl, A. E., & Kennard, R. W. Ridge Regression: Application to Nonorthogonal Problems. Technometrics, 1970, 69-82. (b).

Johnson, J. Econometric Methods. New York: McGraw-Hill Book Co., 1963.

Kerridge, D. Error of Prediction in Multiple Regression with Stochastic Regressor Variables. Technometrics, 1967, 9, 309-311.

Klingler, D. E. Improvement of prediction for nonorthogonal problems: A cross validation study of ridge regression. Doctoral Dissertation, University of Colorado, 1975.

Laughlin, J. E. Comment on "Estimating Coefficients in Linear Model: It Don't Make No Nevermind". Psychological Bulletin, 1978, 85, 247-253.

Lawshe, C. H., & Schucker, R. E. The relative efficiency of four test weighting methods in multiple prediction. Educational and Psychological Measurement, 1959, 19, 103-114.

Lord, F. M., & Novick, M. R. Statistical Theories of Mental Test Scores. Reading Massachusetts: Addison-Wesley Publishing Company, 1968.

Malinvaud, E. Statistical Methods of Econometrics. Amsterdam: North-Holland Publishing Co., 1966.

Marquardt, D. W. Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation. Technometrics, 1970, 12, 591-612.

Marquardt, D. W., & Snee, R. D. Ridge Regression in Practice. American Statistician, 1975, 29, 3-20.

Narula, S. C. Predictive Mean Square Error and Stochastic Regressor Variables. Applied Statistics, 1974, 23, 11-17.

Nunnally, J. C. Psychometric Theory. New York: McGraw-Hill Book Company, 1967.

Obenchain, R. L. Ridge Analysis Following a Preliminary Test of the Shrunken Hypothesis. Technometrics, 1975, 14, 431-442.

Perloff, R. Using Trend Fitting Predictor Weights to Improve Cross-Validation. Doctoral Dissertation, The Ohio State University, 1951.

Pruzek, R. M., & Frederick, B. C. Weighting Predictors in Linear Models: Alternatives to Least Squares and Limitations of Equal Weights. Psychological Bulletin, 1978, 85, 254-266.

Rao, C. R. Linear Statistical Inference and its Application. New York: John Wiley & Sons, 1965.

Sampson, A. R. A tale of two regressions. Journal of the American Statistical Association, 1974, 69, 682-689.

Schmidt, F. L. The relative efficiency of regression and simple unit predictor weights in applied differential psychology. Educational and Psychological Measurement, 1971, 31, 699-714.

Schneeweiss, H. Consistent estimation of a regression with errors in the variables. Metrika, 1976, 23, 101-115.

Stein, C. M. Multiple regression. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann, Contributions to Probability and Statistics. Essays in Honor of Harold Hottelling. Stanford, California: Stanford University Press, 1960.

Wainer, H. Estimating Coefficients in Linear Model: It don't make no nevermind. Psychological Bulletin, 1976, 83, 213-217.

Wainer, H. On the Sensitivity of Regression and Regressors. Psychological Bulletin, 1978, 85, 267-273.

Wainer, H., & Thissen, D. Three steps towards robust regression. Psychometrika, 1976, 41, 9-34.

Warren, R. D., White, J. K., & Fuller, W. A. An Error-In-Variables Analysis of Managerial Role Performance. Journal of the American Statistical Association, 1974, 69, 886-893.

Wesman, A. G., & Bennett, G. K. Multiple regression vs. simple addition of scores in prediction of college grades. Educational and Psychological Measurement, 1959, 19, 243-246.

Wilks, S. S. Weighting system for linear functions of correlated variables when there is no dependent variable. Psychometrika, 1938, 3, 23-40.

Wolins, L. The use of multiple regression procedures when the predictor variables are psychological test. Educational Psychological Measurement, 1967, 27, 821-827.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Leroy Wolins and Dr. William J. Kennedy for their encouragement and guidance throughout this research. Dr. C. P. Han offered many helpful suggestions for solving multivariate problems.

Special thanks go to my mother for her constant encouragement.

My greatest thanks go to my wife, Miriam, for her patience. She was a working mother in El Salvador while this research was conducted in Ames.